(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 August 2002 (01.08.2002)

PCT

(10) International Publication Number
WO 02/059330 A2

(51) International Patent Classification[7]: C12N 15/67

(21) International Application Number: PCT/DK02/00058

(22) International Filing Date: 25 January 2002 (25.01.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PA 2001 00130 25 January 2001 (25.01.2001) DK
60/300,865 27 June 2001 (27.06.2001) US

(71) Applicant *(for all designated States except US)*: **EVOLVA BIOTECH A/S** [DK/DK]; Fruebjergvej 3, DK-2100 Copenhagen Ø (DK).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **GOLDSMITH, Neil** [GB/GB]; 17 Southfield Road, Oxford OX4 1NX (GB). **SØRENSEN, Alexandra, M., P., SantAna** [PT/DK]; Østre Paradisvej 5D, st.th., DK-2840 Holte (DK). **NIELSEN, Søren, V., S.** [DK/DK]; Høveltsvangvej 72, DK-3450 Allerød (DK).

(74) Agent: **HØIBERG APS**; Store Kongensgade 59 B, DK-1264 Copenhagen K. (DK).

(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), DE (utility model), DK (utility model), DM, DZ, EC, EE (utility model), ES, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/059330 A2

(54) Title: ARTIFICIAL CHROMOSOMES COMPRISING CONCATEMERS OF EXPRESSIBLE NUCLEOTIDE SEQUENCES

(57) Abstract: In the present invention is disclosed the use of artificial chromosomes for the coordinated and controllable expression of large numbers of heterologous genes in a single host cell. In particular, the invention relates to an artificial chromosome comprising at least two co-ordinatedly expressible nucleotide sequences, an artificial chromosome comprising at least two expression cassettes and a host cell comprising at least one of these artificial chromosomes as well as to a host cell comprising at least three different artificial chromosomes.

Artificial chromosomes comprising concatemers for expressible nucleotide sequences.

This application is a nonprovisional of U.S. provisional application Serial No. 60/300,865 filed 27 June 2001, which is hereby incorporated by reference in its entirety. The application claims priority from Danish patent application number PA 2001 00130 filed 25 January 2001, which is hereby incorporated by reference in its entirety. All patent and nonpatent references cited in the application, or in the present application, are also hereby incorporated by reference in their entirety.

Field of the invention

In the present invention is disclosed the use of artificial chromosomes for the co-ordinated and controllable expression of large numbers of heterologous genes in a single host cell. In particular, the invention relates to an artificial chromosome comprising at least two co-ordinatedly expressible nucleotide sequences, an artificial chromosome comprising at least two expression cassettes and a host cell comprising at least one of these artificial chromosomes as well as to a host cell comprising at least three different artificial chromosomes.

Prior art

An artificial chromosome is a vector based on functional entities derived from a natural chromosome that can replicate and be stably maintained in a cell.

Artificial chromosomes are man-made linear or circular DNA molecules constructed from essential cis-acting DNA sequence elements that are responsible for the proper replication and partitioning of natural chromosomes (see Murray et al. Nature 301:189-193 (1983)). These essential elements are: (1) Autonomous Replication Sequences (ARS) (have properties of replication origins, which are the sites for initiation of DNA replication). (2) Centromeres (site of kinetochore assemble and responsible for proper distribution of replicated chromosomes at meiosis and mitosis), and (3) Telomeres (specialised structures at the ends of linear chromosomes that function to stabilise the ends and facilitate the complete replication of the extreme termini of the DNA molecule).

Artificial chromosomes have been constructed in yeast using the three cloned essential chromosomal elements. Murray et al., Nature 305:189-193 (1983), disclose a cloning system based on the In vitro construction of linear DNA molecules that can be transformed into yeast, where they are maintained as artificial chromosomes. These artificial yeast chromosomes contain cloned genes, replicators, centromeres and telomeres but have impaired centromeric function in short (less than 20 kb) artificial chromosomes. Another Yeast artificial chromosome, called a functional minichromosome is disclosed in US 4,464,472 (Carbon et al).

Artificial chromosomes have been constructed for a number of species and methods have been developed to generalise the design of artificial chromosomes for other species.

US 5,270,201 (Richards et al) describe an artificial chromosome vector which is especially adapted for insertion into plant cells such as Arabidopsis thaliana.

Hamilton et al (US 5,977,439) have developed a so-called BIBAC vector for Agrobacterium based transformation of plant cells. The BIBAC vector is based on a Bacterial Artificial Chromosome (BAC) and a binary vector (BIN). The BIBAC vector allows construction of plant genomic libraries with large DNA inserts that can be introduced into plants by transformation mediated by Agrobacterium.

Artificial chromosomes based on Baculovirus may be used as artificial chromosomes in insects such as Lepidoptera including butterflies and moths (US 6,090,584 (Iatrou et al)).

Recently, methods for preparation of mammalian artificial chromosomes have also been developed (US 6,133,503 (Scheffler) and US 6,077,697 (Hadlaczky et al)) and it must be envisaged that it becomes possible to design suitable artificial chromosomes for any desired species.

Artificial chromosomes can be regarded as giant vectors adapted to stably maintain in the host cell, large nucleotide sequences. Artificial chromosomes have been used as libraries of nucleotide sequences, for gene therapy, especially gene therapy

involving the simultaneous expression of an entire metabolic pathway. Apart from this, artificial chromosomes may be used as information storage vehicles, for analysis and study of centromere function. Known artificial chromosomes include chromosomes comprising up to 1000 megabases.

Another application (WO 99/67374) of artificial chromosomes is an application, whereby one transfers the ability to produce a secondary metabolite from an actinomycete that is the original producer of the natural product, to a different production host that has desirable characteristics. The application involves the construction of a segment of the chromosome of the original producer in an artificial chromosome that can be stably maintained in a suitable production host.

Artificial chromosomes have not been used for the co-ordinated and controlled expression of a number of different genes and artificial chromosomes have not been used in the evolution of novel biochemical pathways.

Summary of the invention

In a first aspect the invention relates to an artificial chromosome comprising at least one nucleotide concatemer, the concatemer comprising in the 5'→3' direction a cassette of nucleotide sequence of the general formula

$[rs_2\text{-SP-PR-X-TR-SP-}rs_1]_n$

wherein

rs_1 and rs_2 together denote a restriction site,

SP denotes a spacer of at least two nucleotide bases,

PR denotes a promoter, capable of functioning in a cell,

X denotes an expressible nucleotide sequence,

TR denotes a terminator, and

SP denotes a spacer of at least two nucleotide bases, and

$n \geq 2$.

Due to the highly ordered structure of the concatemer the assembly of the concatemer is easily performed, especially when the restriction site comprises sticky ends having a pre-determined nucleotide sequence. The expressible nucleotide sequences may conveniently arise from a cDNA library obtained from one or more

expression states, wherein the cDNA clones have been inserted into expression cassettes. Following excision of the expression cassettes from the vector comprising the construct in the cDNA library, the multitude of constructs may be concatenated and inserted into an "empty" artificial chromosome for subsequent transformation into a host cell.

The artificial chromosome according to the invention may comprise a selection of expressible nucleotide sequences from just one expression state and can thus be assembled from one library representing this expression state or it may comprise cassettes from a number of different expression states. The variation among and between cassettes in the artificial chromosome may be such as to minimise the chance of cross over as the host cell undergoes cell division such as through minimising the level of repeat sequences occurring in any one concatemer, since it is not an object of this embodiment of the invention to obtain inter- or intrachromosomal recombination of the artificial chromosomes. Nor is it an object to obtain recombination with the host genome or an episome of the host cells.

One advantage of the structure of the concatemer is that it can be recovered from the host cell and by subsequent digestion with a restriction enzyme specific for the rs_1-rs_2 restriction site. The building blocks of the concatemers may thus be disassembled and reassembled at any point.

The cassettes of the concatemer may be joined head to tail or head to head or tail to tail, which does not affect expression of the expressible nucleotide sequences because each expressible nucleotide sequence is under the control of it's own promoter. This is due to the fact that most restriction enzymes leave two identical overhangs, which may combine in either order at the same frequency.

In a second aspect the invention relates to an artificial chromosome comprising at least a first and a second expressible nucleotide sequence under the control of a controllable promoter, the promoter of the first expressible nucleotide sequence being controllable independently from the promoter of the other expressible nucleotide sequence.

By having two or more expressible nucleotide sequences located on the same artificial chromosome under the control of different promoters, the expression state of a cell comprising the artificial chromosome can be manipulated in a co-ordinated way through regulation of the two or more different promoters. The artificial chromosomes are especially useful in the evolution of novel biochemical pathways, where genes from multiple expression states (e.g. from multiple species) are combined in one host cell. The single genes may be inserted under the control of different promoters. Preferably one artificial chromosome comprises a unique combination of promoters and genes. By having several artificial chromosomes inserted into a number of cells, in principle any combination of sub-sets of genes may be turned on or off in a population of cells by having random combinations of genes and promoters represented. Furthermore, by up and down regulation of specific promoters, different sub-sets of genes may be turned on and off in a co-ordinated way and numerous combinations of expressed genes may be obtained in just one cell. Furthermore, in biochemical pathway evolution, chances are great that lethal genes are inserted into the host cell. Through down regulation of different promoters, those controlling the lethal genes may be switched off allowing evolution of biochemical pathways from the remaining non-lethal genes.

In a further aspect the invention relates to a host cell comprising at least one artificial chromosome comprising at least a first and a second expressible nucleotide sequence under the control of a controllable promoter, the promoter of the first expressible nucleotide sequence being controllable independently from the promoter of the other expressible nucleotide sequence.

Such host cells are ideal candidates for the evolution of novel biochemical pathways leading possibly to novel metabolites, such as drug candidates. The expression state of the transgenic cell may be changed in a co-ordinated way through up or down regulation of one or more controllable promoters. As explained above identical promoters preferably regulates a subset of expressible nucleotide sequences allowing the co-ordinated expression of sub-sets of genes. In a population of cells according to the invention, multiple combinations of genes may be co-ordinatedly expressed in this way.

In another aspect the invention relates to a host cell comprising at least two artificial chromosomes containing a concatemer each. By having at least two artificial chromosomes in one cell, evolution can be performed using techniques such as traditional breeding.

In a still further aspect the invention relates to a host cell comprising at least three artificial chromosomes, wherein the three chromosomes are different. More preferably the invention relates to a host cell comprising at least four artificial chromosomes, wherein the four chromosomes are different.

By having at least three different artificial chromosomes in one cell, a very high number of foreign genes can be inserted and maintained in the host cell. The host cell may either be used as a library cell for information storage purposes or the artificial chromosomes may comprise expressible gene sequences for gene therapy, for production of proteins for production of compounds requiring the expression of a high number of genes and/or for evolution of novel biochemical pathways.

Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of skill in the art to which this invention belongs. All patents and publications referred to herein are incorporated by reference.

As used herein, a mammalian artificial chromosome [MAC] is a piece of DNA that can stably replicate and segregate alongside endogenous chromosomes. It has the capacity to accommodate and express heterologous genes inserted therein. It is referred to as a mammalian artificial chromosome because it includes an active mammalian centromere. Plant artificial chromosomes and an insect artificial chromosomes refer to chromosomes that include plant and insect centromeres, respectively. A human artificial chromosome [HAC] refers to chromosomes that include human centromeres, BUGACs refer to artificial insect chromosomes, and AVACs refer to avian artificial chromosomes. A yeast artificial chromosome (YAC) refers to chromosomes that includes centromere being functional in yeast, such as a yeast centromere.

As used herein, stable maintenance of chromosomes, occurs when at least about 85%, preferably 90%, more preferably 95%, of the cells retain the chromosome. Stability is measured in the presence of selective agent. Preferably these chromosomes are also maintained in the absence of a selective agent. Stable chromosomes also retain their structure during cell culturing, suffering neither intrachromosomal nor interchromosomal rearrangements.

As used herein, growth under selective conditions, means growth of a cell under conditions that require expression of a selectable marker for survival.

By a controllable promoter is meant a promoter, which can be controlled through external manipulations such as addition or removal of a compound from the surroundings of the cell, change of physical conditions, etc.

Co-ordinated expression refers to the expression of a sub-set of genes which are induced or repressed by the same external stimulus or stimuli.

Restriction site

For the purposes of the present invention the abbreviation RSn (n=1,2,3, etc) is used to designate a nucleotide sequence comprising a restriction site. A restriction site is defined by a recognition sequence and a cleavage site. The cleavage site may be located within or outside the recognition sequence. The abbreviation "rs_1" or "rs_2" is used to designate the two ends of a restriction site after cleavage. The sequence "rs_1-rs_2" together designate a complete restriction site.

The cleavage site of a restriction site may leave a double stranded polynucleotide sequence with either blunt or sticky ends. Thus, "rs_1" or "rs_2" may designate either a blunt or a sticky end.

In the notation used throughout the present invention, formulae like:
RS1-RS2-SP-PR-X-TR-SP-RS2-RS1
should be interpreted to mean that the individual sequences follow in the order specified. This does not exclude that part of the recognition sequence of e.g. RS2 overlap with the spacer sequence, but it is a strict requirement that all the items

except RS1 and RS1' are functional and remain functional after cleavage and re-assemblage. Furthermore the formulae do not exclude the possibility of having additional sequences inserted between the listed items. For example introns can be inserted as described in the invention below and further spacer sequences can be inserted between RS1 and RS2 and between TR and RS2. Important is that the sequences remain functional.

Furthermore, when reference is made to the size of the restriction site and/or to specific bases within it, only the bases in the recognition site are referred to.

Expression state

An expression state is a state in any specific tissue of any individual organism at any one time. Any change in conditions leading to changes in gene expression leads to another expression state. Different expression states are found in different individuals, in different species but they may also be found in different organs in the same species or individual, and in different tissue types in the same species or individual. Different expression states may also be obtained in the same organ or tissue in any one species or individual by exposing the tissues or organs to different environmental conditions comprising but not limited to changes in age, disease, infection, drought, humidity, salinity, exposure to xenobiotics, physiological effectors, temperature, pressure, pH, light, gaseous environment, chemicals such as toxins.

Brief description of the drawings

Fig. 1 shows a flow chart of the steps leading from an expression state to incorporation of the expressible nucleotide sequences in an entry library (a nucleotide library according to the invention).

Fig. 2 shows a flow chart of the steps leading from an entry library comprising expressible nucleotide sequences to evolvable artificial chromosomes (EVAC) transformed into an appropriate host cell. Fig. 2a shows one way of producing the EVACs which includes concatenation, size selection and insertion into an artificial chromosome vector. Fig. 2b shows a one step procedure for concatenation and ligation of vector arms to obtain EVACs.

Fig. 3 shows a model entry vector. MCS is a multi cloning site for inserting expressible nucleotide sequences. Amp R is the gene for ampicillin resistance. Col E is the origin of replication in E. coli. R1 and R2 are restriction enzyme recognition sites.

Fig. 4 shows an example of an entry vector according to the invention, EVE4. MET25 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer 1 and spacer 2 are constituted by a few nucleotides deriving from the multiple cloning site, MCS, SrfI and AscI are restriction enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 1.

Fig 5 shows an example of an entry vector according to the invention, EVE5. CUP1 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer 1 and spacer 2 are constituted by a few nucleotides deriving from the multiple cloning site, MCS, SrfI and AscI are restriction enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 2.

Fig 6 shows an example of an entry vector according to the invention, EVE8. CUP1 is a promoter, ADH1 is a terminator, f1 is an origin of replication for filamentous phages, e.g. M13. Spacer3 is a 550 bp fragment of lambda phage DNA. Spacer4 is a ARS1 sequence from yeast. SrfI and AscI are restriction enzyme recognition sites. Other abbreviations, see Fig. 3. The sequence of the vector is set forth in SEQ ID NO 3.

Fig. 7 shows a vector (pYAC4-AscI) for providing arms for an evolvable artificial chromosome (EVAC) into which a concatemer according to the invention can be cloned. TRP1, URA3, and HIS3 are yeast auxotrophic marker genes, and AmpR is an E. coli antibiotic marker gene. CEN4 is a centromere and TEL are telomeres. ARS1 and PMB1 allow replication in yeast and E. coli respectively. BamH I and Asc I are restriction enzyme recognition sites. The nucleotide sequence of the vector is set forth in SEQ ID NO 4.

Fig 8. shows the general concatenation strategy. On the left is shown a circular entry vector with restriction sites, spacers, promoter, expressible nucleotide sequence and terminator. These are excised and ligated randomly.

Lane	F/Y
1	100/1
2	50/1
3	20/1
4	10/1
5	5/1
6	2/1
7	1/1
8	1/2
9	1/5

Legend: Lane M: molecular weight marker, λ–phage DNA digested w. Pst1. Lanes 1-9, concatenation reactions. Ratio of fragments to yac-arms(F/Y) as in table.

Fig 9a and 9b. illustrates the integration of concatenation with synthesis of evolvable artificial chromosomes and how concatemer size can be controlled by controlling the ratio of vector arms to expression cassettes, as described in example 7.

Fig 10. Library of EVAC transformed population shown under 4 different growth conditions. Coloured phenotypes can be readily detected upon induction of the Met25 and/or the CapI promoter.

Fig 11. EVAC gel Legend: PFGE of EVAC containing clones :
Lanes. a: Yeast DNA PFGE markers(strain YNN295), b: lambda ladder, c: non-transformed host yeast, 1 – 9 : EVAC containing clones. EVACs in size range 1400-1600 kb. Lane 2 shows a clone containing 2 EVACs sized ~1500 kb and ~550 kb respectively. The 550kb EVAC is comigrating with the 564kb yeast chromosome and is resulting in an increased intensity of the band at 564 kb relative to the other bands in the lane. Arrows point up to EVAC bands.

Detailed description

In describing the artificial chromosomes of this invention, the individual components will first be considered: Namely the functional element of which the artificial

chromosome is composed; and other genes which contribute properties to transformed cells.

Centromere

The centromere is the junction between the two arms of a chromosome to which the spindle fibers attach, either directly or indirectly, during mitosis and meiosis. Thus, the centromere acts to orient the chromosome during cell splitting, so that the two copies of the chromosome are directed to opposite poles of the cell prior to splitting into two progeny. The centromere also acts as a binding site for binding the chromosome to the spindle, thus ensuring that each daughter cell receives a copy of the chromosome.

Each of the chromosomes of a eukaryote may have a centromere of different composition. For the most part, the centromeres will be relatively small, usually smaller than about 2kbp, usually less than about 1.6kbp and may function with as few as 0.2kbp, more usually as few as 0.5kbp. For the most part, the centromere segment does not have long repetitive segments as observed with heterochromatin.

The centromere may be obtained from any eukaryotic host. Eukaryotic hosts include plants, insects, molds, fungi, mammals and the like. Of particular interest are plants, particularly food crops, fruit trees, and wood trees; fungi, such as mushrooms, yeast; mammals, such as domestic animals and humans; and birds, such as domestic poultry.

There are a number of different ways to obtain centromeres. Initially, the centromere will normally be obtained from a host chromosome. Desirably, the host chromosome has been mapped so as to establish an area which functions as the centromere and is bordered by restriction sites. The area defined as the centromere frequently can be detected by the substantial absence of recombination events in the vicinity of the centromere. By appropriate mapping, one can define structural genes on opposite sides of the centromere and restriction sites which allow for cleavage of the chromosome to produce a segment including at least one structural gene and preferably both structural genes. The structural genes serve as markers, since the expression of the structural genes in a clone requires the presence of the

centromere.

The fragments will generally be less than ten percent in number of base pairs of the chromosome from which the centromere containing fragment was derived. Fragments may then be formed by restriction enzyme cleavage. The fragments may be inserted into a shuttle vector containing a prokaryotic replication site and a eukaryotic chromosomal replicator. By transforming a prokaryote auxotrophic mutant which is complemented by at least one of the structural genes adjacent the centromere one can select for clones having a high probability of having the centromere DNA sequence. Selective medium will permit selection of the transformed clones.

The eukaryotic fragments inserted into the shuttle vector are then excised at the restriction sites; the resulting mixture of eukaryotic segments will have a greatly enhanced concentration of centromere containing segments. The mixture of DNA fragments may now be inserted in the same shuttle vector or a different vector having a replicating site for the host to be transformed, which may or may not be the same host from which the centromere was obtained. Desirably, the host should be an auxotroph for one of the structural genes associated with the centromere to allow for rapid selection of host transformed with the hybrid DNA containing the structural gene. By cultivating the host through a number of generations, transformed cells having plasmid lacking the centromere will be unstable and reject the plasmid. Those cells which retain the markers and are prototrophic in the marker will have plasmids containing the centromere. Therefore, it is not necessary to employ an auxotrophic mutant, it will be sufficient to employ a phenotypic marker, particularly one allowing for selection.

The plasmids are isolated from the cells and by employing overlap hydridization, the DNA sequence providing the centromere function is identified. The centromere may then be isolated substantially free of the genes immediately adjacent the centromere in the chromosome from which the centromere was derived. In this way, one can have a DNA segment which provides the centromere function and can be bonded to a wide variety of structural genes, operators, binding sites, regulating genes, or the like, in addition to the one or more replicating sites.

Once the centromere segment has been isolated, the segment may be sequenced and synthesized.

Replication Site

In order to have stable mitotic maintenance, a replication site in combination with the centromere segment is necessary. The replication site is the DNA sequence which is recognised by the enzymes and proteins involved in replication of the DNA duplex. The replication site can be initially obtained by genomic cloning. The chromosomes of the host can be fragmented either mechanically or preferably by restriction enzymes. The fragments may then be inserted into an appropriate vector, which may or may not have one or more genetic markers. Particularly, the vector should lack a replication site which would allow for replication in the eukaryotic host to be transformed.

After transformation and passage through a number of generations, one can select for the presence of the marker. Only those cells containing a DNA fragment having a replication site will be able to retain the plasmid to any detectible degree. The cells may then be harvested, lysed, and the plasmid isolated. The inserted DNA fragment may be excised and used for introduction of the replication site in combination with the centromere. The replication site will hereinafter be referred to as an autonomously replicating segment, ARS.

Where an autonomously replicating segment is known to be associated with a structural gene, the structural gene may be employed as a marker. By transforming hosts which are auxotrophic for the product expressed by the marker, one can select for transformed cells which are able to grow in a selective medium. Only those cells having the combination of the ARS and marker will survive in the selective medium.

Once the ARS has been isolated as part of a larger fragment, the fragment may be reduced in size, employing endo- or exonucleases, capable of cleavage or processive oligonucleotide removal. The resulting fragments may be inserted in an appropriate vector and used for transformation. Once again, only those cells which

are transformed with a functional ARS will be able to retain the plasmid in selective medium. If the vector includes a centromere, nonselective medium may be employed, since a plasmid containing only the ARS and not the centromere is mitotically unstable.

The ARS fragment may or may not be joined to the native genes on opposite sides of the ARS when combined with the centromere to form the artificial chromosome. When the ARS employed is free of the native functional genes, it will normally be less than about 1kbp, usually less than about 0.5kbp and may be as small as 0.2 kbp.

As part of the artificial chromosome, the ARS may or may not be derived from the same host as the centromere was derived from, nor from the same cell source as the host cell to be transformed by the artificial chromosome.

Telomeres

Telomeres, the last chromosomal element in lower eukaryotes to be cloned, are thought to be involved in the priming of DNA replication at the chromosome end. This is because conventional DNA polymerases are template dependent, synthesise DNA in the 5' to 3' direction, and require an oligonucleotide primer to donate a 3' OH group. When this primer is removed, unreplicated single-stranded gaps arise; most of these gaps can be filled in by priming from 3' OH groups donated by newly replicated strands located at the 5' end of the gap. However, the unreplicated gaps which lie next to the extreme 5' end of the DNA duplex cannot be primed in this manner. Consequently, telomeres must provide an alternative priming mechanism.

Telomeres are also responsible for the stability of chromosomal termini. Telomeres act as "caps," suppressing the recombinogenic properties of free, unmodified DNA ends. This reduces the formation of damaged and rearranged chromosomes which arise as a consequence of recombination-mediated chromosome fusion events.

Telomeres may also contribute to the establishment or maintenance of intranuclear chromatin organization through their association with the nuclear envelope.

Telomeric or telomeric-like DNA sequences have been cloned from several lower eukaryotic organisms, principally protozoans and yeast. The ends of the Tetrahymena linear DNA plasmid have been shown to function like a telomere on linear plasmids in Saccharomyces cerevisiae (see Szostak, J. W., Cold Spring Harbor Symp. Quant. Biol. 47:1187-1194 (1983)). A telomere from the flagellate Trypanosoma has been cloned (see, for example, Blackburn et al., Cell 36:447-457 (1984). A yeast telomeric sequence has been identified (see, for example, Shampay et al., Nature 310:154-157 (1984)).

Telomeres have also been identified in mammalian chromosomes for use in Mammalian Artificial Chromosomes (US 6,133,503)

Artificial chromosome

The artificial chromosome is a combination of a DNA segment comprising a centromeric function, a replicating site (ARS), and telomeres, and one or more genes, including regulatory genes and structural genes, which are to be expressed by the transformed host cell.

Transformation can be achieved by using calcium shock, by exposing host cell spheroplasts to the plasmid DNA under conditions favoring spheroplast fusion and then plating the spheroplast in regeneration agar selecting for the desired phenotype; or other conventional techniques.

The transformed host cells may then be grown on selective or nonselective medium. While the artificial chromosome has mitotic stability, it is well established that aneuploid cells will frequently lose one of the chromosomes. Since the artificial chromosome in nonselective medium will not be necessary for viability, loss of the artificial chromosome will not adversely affect the viability of the resulting "wild type" of cell. Therefore, it will usually be desirable to have a marker on the artificial chromosome which provides for selective pressure for the transformed host cells.

The nature of the marker may be varied widely providing for resistance to a cell growth inhibitor; complementation of an auxotrophic mutation in the transformed host; morphologic change; or the like.

The host cells according to this invention may comprise one or several artificial chromosomes. When the cells comprise more than one artificial chromosome, their presence may be ensured by using a common marker present on all chromosomes. However it may be more advantageous to provide each artificial chromosome with a unique marker and select for cells having markers corresponding to the artificial chromosomes, that they are supposed to contain.

Each cell according to the invention may comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more artificial chromosomes. Each of these chromosomes may be laid out as defined in the claims.

The chromosomes may be maintained in haploid or diploid host cells. Haploid cells may be combined to form diploid cells, which undergo meiosis. Upon meiosis new combinations of chromosomes may be obtained in the offspring.

Origin of expressible nucleotide sequences

The expressible nucleotide sequences that can be inserted into the vectors, concatemers, and cells according to this invention encompass any type of nucleotide such as RNA, DNA. Such a nucleotide sequence could be obtained e.g. from cDNA, which by its nature is expressible. But it is also possible to use sequences of genomic DNA, coding for specific genes. Preferably, the expressible nucleotide sequences correspond to full length genes such as substantially full length cDNA, but nucleotide sequences coding for shorter peptides than the original full length mRNAs may also be used. Shorter peptides may still retain the catalytic activity similar to that of the native proteins.

Another way to obtain expressible nucleotide sequences is through chemical synthesis of nucleotide sequences coding for known peptide or protein sequences. Thus the expressible DNA sequences does not have to be a naturally occurring sequence, although it may be preferable for practical purposes to primarily use naturally occurring nucleotide sequences. Whether the DNA is single or double stranded will depend on the vector system used.

In most cases the orientation with respect to the promoter of an expressible nucleotide sequence will be such that the coding strand is transcribed into a proper mRNA. It is however conceivable that the sequence may be reversed generating an antisense transcript In order to block expression of a specific gene.

Cassettes

An important aspect of the invention concerns a cassette of nucleotides in a highly ordered sequence, the cassette having the general formula in 5'→3' direction:

[RS1-RS2-SP-PR-CS-TR-SP-RS2'-RS1']

wherein RS1 and RS1' denote restriction sites, RS2 and RS2' denote restriction sites different from RS1 and RS1', SP individually denotes a spacer sequence of at least two nucleotides, PR denotes a promoter, CS denotes a cloning site, and TR denotes a terminator.

It is an advantage to have two different restriction sites flanking both sides of the expression construct. By treating the primary vectors with restriction enzymes cleaving both restriction sites, the expression construct and the primary vector will be left with two non-compatible ends. This facilitates a concatenation process, since the empty vectors do not participate in the concatenation of expression cassettes .

Restriction sites

In principle, any restriction site, for which a restriction enzyme is known can be used. These include the restriction enzymes generally known and used in the field of molecular biology such as those described in Sambrook, Fritsch, Maniatis, "A laboratory Manual", 2nd edition. Cold Spring Harbor Laboratory Press, 1989.

The restriction site recognition sequences preferably are of a substantial length, so that the likelihood of occurrence of an identical restriction site within the cloned oligonucleotide is minimised. Thus the first restriction site may comprise at least 6 bases, but more preferably the recognition sequence comprises at least 7 or 8 bases. Restriction sites having 7 or more non N bases in the recognition sequence are generally known as "rare restriction sites" (see example 6). However, the recognition sequence may also be at least 10 bases, such as at least 15 bases, for

example at least 16 bases, such as at least 17 bases, for example at least 18 bases, such as at least 18 bases, for example at least 19 bases, for example at least 20 bases, such as at least 21 bases, for example at least 22 bases, such as at least 23 bases, for example at least 25 bases, such as at least 30 bases, for example at least 35 bases, such as at least 40 bases, for example at least 45 bases, such as at least 50 bases.

Preferably the first restriction site RS1 and RS1' is recognised by a restriction enzyme generating blunt ends of the double stranded nucleotide sequences. By generating blunt ends at this site, the risk that the vector participates in a subsequent concatenation is greatly reduced. The first restriction site may also give rise to sticky ends, but these are then preferably non-compatible with the sticky ends resulting from the second restriction site, RS2 and RS2' and with the sticky ends in the AC.

According to a preferred embodiment of the invention, the second restriction site, RS2 and RS2' comprises a rare restriction site. Thus, the longer the recognition sequence of the rare restriction site the more rare it is and the less likely is it that the restriction enzyme recognising it will cleave the nucleotide sequence at other – undesired – positions.

The rare restriction site may furthermore serve as a PCR priming site. Thereby it is possible to copy the cassettes via PCR techniques and thus indirectly "excise" the cassettes from a vector.

Spacer sequence

The spacer sequence located between the RS2 and the PR sequence is preferably a non-transcribed spacer sequence. The purpose of the spacer sequence(s) is to minimise recombination between different concatemers present in the same cell or between cassettes present in the same concatemer, but it may also serve the purpose of making the nucleotide sequences in the cassettes more "host" like. A further purpose of the spacer sequence is to reduce the occurrence of hairpin formation between adjacent palindromic sequences, which may occur when cassettes are assembled head to head or tail to tail. Spacer sequences may also be convenient

for introducing short conserved nucleotide sequences that may serve e.g. as PCR primer sites or as target for hybridization to e.g. nucleic acid or PNA or LNA probes allowing affinity purification of cassettes.

The cassette may also optionally comprise another spacer sequence of at least two nucleotides between TR and RS2. When cassettes are cut out from a vector and concatenated into concatemers of cassettes, the spacer sequences together ensure that there is a certain distance between two successive identical promoter and/or terminator sequences. This distance may comprise at least 50 bases, such as at least 60 bases, for example at least 75 bases, such as at least 100 bases, for example at least 150 bases, such as at least 200 bases, for example at least 250 bases, such as at least 300 bases, for example at least 400 bases, for example at least 500 bases, such as at least 750 bases, for example at least 1000 bases, such as at least 1100 bases, for example at least 1200 bases, such as at least 1300 bases, for example at least 1400 bases, such as at least 1500 bases, for example at least 1600 bases, such as at least 1700 bases, for example at least 1800 bases, such as at least 1900 bases, for example at least 2000 bases, such as at least 2100 bases, for example at least 2200 bases, such as at least 2300 bases, for example at least 2400 bases, such as at least 2500 bases, for example at least 2600 bases, such as at least 2700 bases, for example at least 2800 bases, such as at least 2900 bases, for example at least 3000 bases, such as at least 3200 bases, for example at least 3500 bases, such as at least 3800 bases, for example at least 4000 bases, such as at least 4500 bases, for example at least 5000 bases, such as at least 6000 bases.

The number of the nucleotides between the spacer located 5' to the PR sequence and the one located 3' to the TR sequence may be any. However, it may be advantageous to ensure that at least one of the spacer sequences comprises between 100 and 2500 bases, preferably between 200 and 2300 bases, more preferably between 300 and 2100 bases, such as between 400 and 1900 bases, more preferably between 500 and 1700 bases, such as between 600 and 1500 bases, more preferably between 700 and 1400 bases.

If the intended host cell is yeast, the spacers present in a concatemer should perferably comprise a combination of a few ARSes with varying lambda phage DNA fragments.

Preferred examples of spacer sequences include but are not limited to: Lamda phage DNA, prokaryotic genomic DNA such as E. coli genomic DNA, ARSes.

Promoter

A promoter is a DNA sequence to which RNA polymerase binds and initiates transcription. The promoter determines the polarity of the transcript by specifying which strand will be transcribed.

- Bacterial promoters normally consist of -35 and -10 (relative to the transcriptional start) consensus sequences which are bound by a specific sigma factor and RNA polymerase.
- Eukaryotic promoters are more complex. Most promoters utilized in expression vectors are transcribed by RNA polymerase II. General transcription factors (GTFs) first bind specific sequences near the transcriptional start and then recruit the binding of RNA polymerase II. In addition to these minimal promoter elements; small sequence elements are recognized specifically by modular DNA-binding / trans-activating proteins (e.g. AP-1, SP-1) which regulate the activity of a given promoter.
- Viral promoters may serve the same function as bacterial and eukaryotic promoters. Upon viral infection of their host, viral promoters direct transcription either by using host transcriptional machinery or by supplying virally encoded enzymes to substitute part of the host machinery. Viral promoters are recognised by the transcriptional machinery of a large number of host organisms and are therefore often used in cloning and expression vectors.

Promoters may furthermore comprise regulatory elements, which are DNA sequence elements which act in conjunction with promoters and bind either repressors (e.g., lacO/ LAC Iq repressor system in E. coli) or inducers (e.g., gal1 /GAL4 Inducer system in yeast). In either case, transcription is virtually "shut off" until the promoter is derepressed or induced, at which point transcription is "turned-on". The choice of promoter in the cassette is primarily dependent on the host organism into which the cassette is intended to be inserted. An important requirement to this end is that the promoter should preferably be capable of

functioning in the host cell, in which the expressible nucleotide sequence is to be expressed.

Preferably the promoter is an externally controllable promoter, such as an inducible promoter and/or a repressible promoter. The promoter may be either controllable (repressible/inducible) by chemicals such as the absence/presence of chemical inducers, e.g. metabolites, substrates, metals, hormones, sugars. The promoter may likewise be controllable by certain physical parameters such as temperature, pH, redox status, growth stage, developmental stage, or the promoter may be inducible/repressible by a synthetic inducer/repressor such as the gal inducer.

In order to avoid unintentional interference with the gene regulation systems of the host cell, and in order to improve controllability of the co-ordinated gene expression the promoter is preferably a synthetic promoter. Suitable promoters are described in US 5,798,227, US 5,667,986. Principles for designing suitable synthetic eukaryotic promoters are disclosed in US 5,559,027, US 5,877,018 or US 6,072,050.

Synthetic inducible eukaryotic promoters for the regulation of transcription of a gene may achieve improved levels of protein expression and lower basal levels of gene expression. Such promoters preferably contain at least two different classes of regulatory elements, usually by modification of a native promoter containing one of the inducible elements by inserting the other of the inducible elements. For example, additional metal responsive elements IR:Es) and/or glucocorticoid responsive elements (GREs) may be provided to native promoters. Additionally, one or more constitutive elements may be functionally disabled to provide the lower basal levels of gene expression.

Preferred examples of promoters include but is not limited to those promoters being induced and/or repressed by any factor selected from the group comprising carbohydrates, e.g. galactose; low inorganic phosphase levels; temperature, e.g. low or high temperature shift; metals or metal ions, e.g. copper ions; hormones, e.g. dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage, e.g. developmental stage; synthetic inducers, e.g. gal inducer. Examples of such promoters include ADH 1, PGK 1, GAP 491, TPI, PYK, ENO, PMA 1, PHO5, GAL 1, GAL 2, GAL 10, MET25, ADH2, MEL 1, CUP 1, HSE, AOX,

MOX, SV40, CaMV, Opaque-2, GRE, ARE, PGK/ARE hybrid, CYC/GRE hybrid, TPI/α2 operator, AOX 1, MOX A.

More preferably, however the promoter is selected from hybrid promoters such as PGK/ARE hybrid, CYC/GRE hybrid or from synthetic promoters. Such promoters can be controlled without interfering too much with the regulation of native genes in the expression host.

Yeast promoters

In the following, examples of known yeast promoters that may be used in conjunction with the present invention are shown. The examples are by no way limiting and only serve to indicate to the skilled practitioner how to select or design promoters that are useful according to the present invention.

Although numerous transcriptional promoters which are functional in yeasts have been described in the literature, only some of them have proved effective for the production of polypeptides by the recombinant route. There may be mentioned in particular the promoters of the PGK genes (3-phosphoglycerate kinase, TDH genes encoding GAPDH (Glyceraldehyde phosphate dehydrogenase), TEF1 genes (Elongation factor 1), MFα1 (α sex pheromone precursor) which are considered as strong constitutive promoters or alternatively the regulatable promoter CYC1 which is repressed in the presence of glucose or PHO5 which can be regulated by thiamine. However, for reasons which are often unexplained, they do not always allow the effective expression of the genes which they control. In this context, it is always advantageous to be able to have new promoters in order to generate new effective host/vector systems. Furthermore, having a choice of effective promoters in a given cell also makes it possible to envisage the production of multiple proteins in this same cell (for example several enzymes of the same metabolic chain) while avoiding the problems of recombination between homologous sequences.

In general, a promoter region is situated in the 5' region of the genes and comprises all the elements allowing the transcription of a DNA fragment placed under their control, in particular:

(1) a so-called minimal promoter region comprising the TATA box and the site of initiation of transcription, which determines the position of the site of initiation as well as the basal level of transcription. In Saccharomyces cerevisiae, the length of the minimal promoter region is relatively variable. Indeed, the exact location of the TATA box varies from one gene to another and may be situated from -40 to -120 nucleotides upstream of the site of the initiation (Chen and Struhl, 1985, EMBO J., 4, 3273-3280)

(2) sequences situated upstream of the TATA box (immediately upstream up to several hundreds of nucleotides) which make it possible to ensure an effective level of transcription either constitutively (relatively constant level of transcription all along the cell cycle, regardless of the conditions of culture) or in a regulatable manner (activation of transcription in the presence of an activator and/or repression in the presence of a repressor). These sequences, may be of several types: activator, inhibitor, enhancer, inducer, repressor and may respond to cellular factors or varied culture conditions.

Examples of such promoters are the ZZA1 and ZZA2 promoters disclosed in US 5,641,661, the EF1-α protein promoter and the ribosomal protein S7 gene promoter disclosed in WO 97/44470,, the COX 4 promoter and two unknown promoters (SEQ ID No: 1 and 2 in the document) disclosed in US 5,952,195. Other useful promoters include the HSP150 promoter disclosed in WO 98/54339 and the SV40 and RSV promoters disclosed in US 4,870,013 as well as the PyK and GAPDH promoters disclosed in EP 0 329 203 A1.

Synthetic yeast promoters

More preferably the invention employs the use of synthetic promoters. Synthetic promoters are often constructed by combining the minimal promoter region of one gene with the upstream regulating sequences of another gene. Enhanced promoter control may be obtained by modifying specific sequences in the upstream regulating sequences, e.g. through substitution or deletion or through inserting multiple copies of specific regulating sequences. One advantage of using synthetic promoters is that they may be controlled without interfering too much with the native promoters of the host cell.

One such synthetic yeast promoter comprises promoters or promoter elements of two different yeast-derived genes, yeast killer toxin leader peptide, and amino terminus of IL-1β (WO 98/54339).

Another example of a yeast synthetic promoter is disclosed in US 5,436,136 (Hinnen et al), which concerns a yeast hybrid promoter including a 5' upstream promoter element comprising upstream activation site(s) of the yeast PHO5 gene and a 3' downstream promoter element of the yeast GAPDH gene starting at nucleotide -300 to -180 and ending at nucleotide -1 of the GAPDH gene.

Another example of a yeast synthetic promoter is disclosed in US 5,089,398 (Rosenberg et al). This disclosure describes a promoter with the general formula -(P.R.(2)-P.R.(1))-
wherein:
P.R.(1) is the promoter region proximal to the coding sequence and having the transcription initiation site, the RNA polymerase binding site, and including the TATA box, the CAAT sequence, as well as translational regulatory signals, e.g., capping sequence, as appropriate;
P.R.(2) is the promoter region joined to the 5'-end of P.R.(1) associated with enhancing the efficiency of transcription of the RNA polymerase binding region;

In US 4,945,046 (Horii et al) discloses a further example of how to design a synthetic yeast promoter. This specific promoter comprises promoter elements derived both from yeast and from a mammal. The hybrid promoter consists essentially of Saccharomyces cerevisiae PHO5 or GAP-DH promoter from which the upstream activation site (UAS) has been deleted and replaced by the early enhancer region derived from SV40 virus.

Cloning site

The cloning site in the cassette in the primary vector should be designed so that any nucleotide sequence can be cloned into it.

The cloning site in the cassette preferably allows directional cloning. Hereby is ensured that transcription in a host cell is performed from the coding strand in the

intended direction and that the translated peptide is identical to the peptide for which the original nucleotide sequence codes.

However according to some embodiments it may be advantageous to insert the sequence in opposite direction. According to these embodiments, so-called antisense constructs may be inserted which prevent functional expression of specific genes involved in specific pathways. Thereby it may become possible to divert metabolic intermediates from a prevalent pathway to another less dominant pathway.

The cloning site in the cassette may comprise multiple cloning sites, generally known as MCS or polylinker sites, which is a synthetic DNA sequence encoding a series of restriction endonuclease recognition sites. These sites are engineered for convenient cloning of DNA into a vector at a specific position and for directional cloning of the insert.

Cloning of cDNA does not have to involve the use of restriction enzymes. Other alternative systems include but are not limited to:

Creator™ Cre-loxP system from Clontech, which uses recombination and loxP sites

use of Lambda attachment sites (att-λ), such as the Gateway™ system from Life Technologies.

Both of these systems are directional.

Terminator

The role of the terminator sequence is to limit transcription to the length of the coding sequence. An optimal terminator sequence is thus one, which is capable of performing this act in the host cell.

In prokaryotes, sequences known as transcriptional terminators signal the RNA polymerase to release the DNA template and stop transcription of the nascent RNA.

In eukaryotes, RNA molecules are transcribed well beyond the end of the mature mRNA molecule. New transcripts are enzymatically cleaved and modified by the

addition of a long sequence of adenylic acid residues known as the poly-A tail. A polyadenylation consensus sequence is located about 10 to 30 bases upstream from the actual cleavage site.

Preferred examples of yeast derived terminator sequences include, but are not limited to: ADN1, CYC1, GPD, ADH1 alcohol dehydrogenase.

Intron

Optionally, the cassette in the vector comprises an intron sequence, which may be located 5' or 3' to the expressible nucleotide sequence. The design and layout of introns is well known in the art. The choice of intron design largely depends on the intended host cell, in which the expressible nucleotide sequence is eventually to be expressed. The effects of having intron sequence in the expression cassettes are those generally associated with intron sequences.

Examples of yeast introns can be found in the literature and in specific databases such as Ares Lab Yeast Intron Database (Version 2.1) as updated on 15 April 2000. Earlier versions of the database as well as extracts of the database have been published in: "Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae." by Spingola M, Grate L, Haussler D, Ares M Jr. (RNA 1999 Feb;5(2):221-34) and "Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast." by Davis CA, Grate L, Spingola M, Ares M Jr, (Nucleic Acids Res 2000 Apr 15;28(8):1700-6).

Primary vectors (entry vectors)

By the term entry vector is meant a vector for storing and amplifying cDNA or other expressible nucleotide sequences using the cassettes according to the present invention. The primary vectors are preferably able to propagate in E. coli or any other suitable standard host cell. It should preferably be amplifiable and amenable to standard normalisation and enrichment procedures.

The primary vector may be of any type of DNA that has the basic requirements of a) being able to replicate itself in at least one suitable host organism and b) allows insertion of foreign DNA which is then replicated together with the vector and c) preferably allows selection of vector molecules that contain insertions of said foreign DNA. In a preferred embodiment the vector is able to replicate in standard hosts like yeasts, and bacteria and it should preferably have a high copy number per host cell. It is also preferred that the vector in addition to a host specific origin of replication, contains an origin of replication for a single stranded virus, such as e.g. the f1 origin for filamentous phages. This will allow the production of single stranded nucleic acid which may be useful for normalisation and enrichment procedures of cloned sequences. A vast number of cloning vectors have been described which are commonly used and references may be given to e.g. Sambrook,J; Fritsch, E.F; and Maniatis T. (1989) Molecular Cloning: A laboratory manual. Cold Spring Harbour Laboratory Press, USA, Netherlands Culture Collection of Bacteria (www.cbs.knaw.nl/NCCB/collection.htm) or Department of Microbial Genetics, National Institute of Genetics, Yata 1111 Mishima Shizuoka 411-8540, Japan (www.shigen.nig.ac.jp/cvector/cvector.html). A few type-examples that are the parents of many popular derivatives are M13mp10, pUC18, Lambda gt 10, and pYAC4. Examples of primary vectors include but are not limited to M13K07, pBR322, pUC18, pUC19, pUC118, pUC119, pSP64, pSP65, pGEM-3, pGEM-3Z, pGEM-3Zf(-), pGEM-4, pGEM-4Z, πAN13, pBluescript II, CHARON 4A, λ^+, CHARON 21A, CHARON 32, CHARON 33, CHARON 34, CHARON 35, CHARON 40, EMBL3A, λ2001, λDASH, λFIX, λgt10, λgt11, λgt18, λgt20, λgt22, λORF8, λZAP/R, pJB8, c2RB, pcos1EMBL

Methods for cloning of cDNA or genomic DNA into a vector are well known in the art. Reference may be given to J. Sambrook, E.F. Fritsch, T. Maniatis: Molecular Cloning, A Laboratory Manual (2nd edition, Cold Spring Harbor Laboratory Press, 1989).

One example of a circular model entry vector is described in Figure 3. The vector, EVE contains the expression cassette, R1-R2-Spacer-Promoter-Multi Cloning Site-Terminator-Spacer-R2-R1. The vector furthermore contains a gene for ampicillin resistance, AmpR, and an origin of replication for E.coli, ColE1.

The entry vectors EVE4, EVE5, and EVE8 shown in Figures 4, 5, and 6. These all contain SrfI as R1 and AscI as R2. Both of these sites are palindromic and are regarded as rare restriction sites having 8 bases in the recognition sequence. The vectors furthermore contain the AmpR ampicillin resistance gene, and the ColE1 origin or replication for E.coli as well as f1, which is an origin of replication for filamentous phages, such as M13. EVE4 (Fig. 4) contains the MET25 promoter and the ADH1 terminator. Spacer 1 and spacer 2 are short sequences deriving from the multiple cloning site, MCS. EVE5 (Fig. 5) contains the CUP1 promoter and the ADH1 terminator. EVE8 (Fig. 6) contains the CUP1 promoter and the ADH1 terminator. The spacers of EVE8 are a 550 bp lambda phage DNA (spacer 3) and an ARS sequence from yeast (spacer 4).

Nucleotide library (entry library)

Methods as well as suitable vectors and host cells for constructing and maintaining a library of nucleotide sequences in a cell are well known in the art. The primary requirement for the library is that is should be possible to store and amplify in it a number of primary vectors (constructs) according to this invention, the vectors (constructs) comprising expressible nucleotide sequences from at least one expression state and wherein at least two vectors (constructs) are different.

One specific example of such a library is the well known and widely employed cDNA libraries. The advantage of the cDNA library is mainly that it contains only DNA sequences corresponding to transcribed messenger RNA in a cell. Suitable methods are also present to purify the isolated mRNA or the synthesised cDNA so that only substantially full-length cDNA is cloned into the library.

Methods for optimisation of the process to yield substantially full length cDNA may comprise size selection, e.g. electrophoresis, chromatography, precipitation or may comprise ways of increasing the likelihood of getting full length cDNAs, e.g. the SMART™ method (Clonetech) or the CapTrap™ method (Stratagene).

Preferably the method for making the nucleotide library comprises obtaining a substantially full length cDNA population comprising a normalised representation of cDNA species. More preferably a substantially full length cDNA population

comprises a normalised representation of cDNA species characteristic of a given expression state.

Normalisation reduces the redundancy of clones representing abundant mRNA species and increases the relative representation of clones from rare mRNA species.

Methods for normalisation of cDNA libraries are well known in the art. Reference may be given to suitable protocols for normalisation such as those described in US 5,763,239 (DIVERSA) and WO 95/08647 and WO 95/11986. and Bonaldo, Lennon, Soares, Genome Research 1996, 6:791-806; Ali, Holloway, Taylor, Plant Mol Biol Reporter, 2000, 18:123-132.

Enrichment methods are used to isolate clones representing mRNA which are characteristic of a particular expression state. A number of variations of the method broadly termed as subtractive hybrisation are known in the art. Reference may be given to Sive, John, Nucleic Acid Res, 1988, 16:10937; Diatchenko, Lau, Campbell et al, PNAS, 1996, 93:6025-6030; Carninci, Shibata, Hayatsu, Genome Res, 2000, 10:1617-30, Bonaldo, Lennon, Soares, Genome Research 1996, 6:791-806; Ali, Holloway, Taylor, Plant Mol Biol Reporter, 2000, 18:123-132. For example, enrichment may be achieved by doing additional rounds of hybridization similar to normalization procedures, using e.g. cDNA from a library of abundant clones or simply a library representing the uninduced state as a driver against a tester library from the induced state. Alternatively mRNA or PCR amplified cDNA derived from the expression state of choice can be used to subtract common sequences from a tester library. The choice of driver and tester population will depend on the nature of target expressible nucleotide sequences in each particular experiment

In the library an expressible nucleotide sequence coding for one peptide is preferably found in different but similar vectors under the control of different promoters. Preferably the library comprises at least three primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of three different promoters. More preferably the library comprises at least four primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of four different promoters. More preferably the library comprises at least

five primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of five different promoters, such as comprises at lest six primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of six different promoters, for example comprises at least seven primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of seven different promoters, for example comprises at least eight primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of eight different promoters, such as comprises at least nine primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of nine different promoters, for example comprises at least ten primary vectors with an expressible nucleotide sequence coding for the same peptide under the control of ten different promoters.

The expressible nucleotide sequence coding for the same peptide preferably comprises essentially the same nucleotide sequence, more preferably the same nucleotide sequence.

By having a library with what may be termed one gene under the control of a number of different promoters in different vectors, it is possible to construct from the nucleotide library an array of combinations of genes and promoters. Preferably, one library comprises a complete or substantially complete combination such as a two dimensional array of genes and promoters, wherein substantially all genes are found under the control of substantially all of a selected number of promoters.

According to another embodiment of the invention the nucleotide library comprises combinations of expressible nucleotide sequences combined in different vectors with different spacer sequences and/or different intron sequences. Thus any one expressible nucleotide sequence may be combined in a two, three, four or five dimensional array with different promoters and/or different spacers and/or different introns and/or different terminators. The two, three, four or five dimensional array may be complete or incomplete, since not all combinations will have to be present.

The library may suitably be maintained in a host cell comprising prokaryotic cells or eukaryotic cells. Preferred prokaryotic host organisms may include but are not

limited to Escherichia coli, Bacillus subtilis, Streptomyces lividans, Streptomyces coelicolor Pseudomonas aeruginosa, Myxococcus xanthus.

Yeast species such as Saccharomyces cerevisiae (budding yeast), Schizosaccharomyces pombe (fission yeast), Pichia pastoris, and Hansenula polymorpha (methylotropic yeasts) may also be used. Filamentous ascomycetes, such as Neurospora crassa and Aspergillus nidulans may also be used. Plant cells such as those derived from Nicotiana and Arabidopsis are preferred. Preferred mammalian host cells include but are not limited to those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS; 293, VERO, HeLa etc (see Kriegler M. in "Gene Transfer and Expression: A Laboratory Manual", New York, Freeman & Co. 1990).

Concatemers

A concatemer is a series of linked units. In the present context a concatemer is used to denote a number of serially linked nucleotide cassettes, wherein at least two of the serially linked nucleotide units comprises a cassette having the basic structure

$[rs_2$-SP-PR-X-TR-SP-$rs_1]$

wherein

rs_1 and rs_2 together denote a restriction site,

SP individually denotes a spacer of at least two nucleotide bases,

PR denotes a promoter, capable of functioning in a cell,

X denotes an expressible nucleotide sequence,

TR denotes a terminator, and

SP individually denotes a spacer of at least two nucleotide bases.

Optionally the cassettes comprise an intron sequence between the promoter and the expressible nucleotide sequence and/or between the terminator and the expressible sequence.

The expressible nucleotide sequence in the cassettes of the concatemer may comprise a DNA sequence selected from the group comprising cDNA and genomic DNA.

According to one aspect of the invention, a concatemer comprises cassettes with expressible nucleotide from different expression states, so that non-naturally occurring combinations or non-native combinations of expressible nucleotide sequences are obtained. These different expression states may represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera. The different species may be from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.

For example, the expressible nucleotide sequences may originate from eukaryots such as mammals such as humans, mice or whale, from reptiles such as snakes crocodiles or turtles, from tunicates such as sea squirts, from lepidoptera such as butterflies and moths, from coelenterates such as jellyfish, anenomes, or corals, from fish such as bony and cartilaginous fish, from plants such as dicots, e.g. coffee, oak or monocots such as grasses, lilies, and orchids; from lower plants such as algae and gingko, from higher fungi such as terrestrial fruiting fungi, from marine actinomycetes. The expressible nucleotide sequences may also originate from protozoans such as malaria or trypanosomes, or from prokaryotes such as E. coli or archaebacteria.. Furthermore, the expressible nucleotide sequences may originate from one or more preferably from more expression states from the species and genera listed in the table below.

Bacteria	Streptomyces , Micromonospora, Norcadia, Actinomadura, Actinoplanes, Streptosporangium, Microbispora, Kitasatosporiam, Azobacterium, Rhizobium, Achromobacterium, Enterobacterium, Brucella, Micrococcus, Lactobacillus, Bacillus (B.t. toxins), Clostridium (toxins), Brevibacterium, Pseudomonas, Aerobacter, Vibrio, Halobacterium, Mycoplasma, Cytophaga, Myxococcus
Fungi	Amanita muscaria (fly agaric, ibotenic acid, muscimol), Psilocybe (psilocybin) Physarium, Fuligo, Mucor, Phytophtora, Rhizopus, Aspergillus, Penicillium (penicillin), Coprinus, Phanerochaete, Acremonium (Cephalosporin), Trochoderma, Helminthosporium, Fusarium, Altemaria, Myrothecium, Saccharomyces
Algae	Digenea simplex (kainic acid, antihelminthic), Laminaria anqustata (laminine, hypotensive)

Lichens	Usnea fasciata (vulpinicacid, antimicrobial; usnic acid, antitumor)
Higher Plants	Artemisia (artemisinin), Coleus (forskolin), Desmodium (K channel agonist), Catharanthus (Vinca alkaloids), Digitalis (cardiac glycosides), Podophyllum (podophyllotoxin), Taxus (taxol), Cephalotaxus (homoharringtonine), Camptotheca (Camptothecin), Camellia sinensis (Tea), Cannabis indica, Cannabis sativa (Hemp), Erythroxylum coca (Coca), Lophophora williamsii (PeyoteMyristica fragrans (Nutmeg), Nicotiana, Papaver somniferum (Opium Poppy), Phalaris arundinacea (Reed canary grass)
Protozoa	Ptychodiscus brevis; Dinoflagellates (brevitoxin, cardiovascular)
Sponges	Microciona prolifera (ectyonin, antimicrobial) Cryptotethya cryta (D-arabino furanosides)
Coelenterata	Portuguese Man o War & other jellyfish and medusoid toxins.
Corals	Pseudoterogonia species (Pseudoteracins, anti-inflammatory), Erythropodium (erythrolides, anti-inflammatory)
Aschelminths	Nematode secretory compounds
Molluscs	Conus toxins, sea slug toxins, cephalapod neurotransmitters, squid inks
Annelida	Lumbriconereis heteropa (nereistoxin, insecticidal)
Arachnids	Dolomedes ("fishing spider" venoms)
Crustacea	Xenobalanus (skin adhesives)
Insects	Epilachna (mexican bean beetle alkaloids)
Spinunculida	Bonellia viridis (bonellin,neuroactive)
Bryozoans	Bugula neritina (bryostatins,anti cancer)
Echinoderms	Crinoid chemistry
Tunicates	Trididemnum solidum (didemnin,anti-tumor and anti-viral; Ecteinascidia turbinata ecteinascidins, anti-tumor)
Vertebrates	Eptatretus stoutii (eptatretin,cardioactive), Trachinus draco (proteinaceous toxins, reduce blood pressure, respiration and reduce heart rate). Dendrobatid frogs

(batrachotoxins, pumiliotoxins, histrionicotoxins, and other polyamines); Snake venom toxins; Orinthorhynohus anatinus (duck-billed platypus venom), modified carotenoids, retinoids and steroids; Avians: histrionicotoxins, modified carotenoids, retinoids and steroids

According to a preferred embodiment of the invention the concatemer comprises at least a first cassette and a second cassette, said first cassette being different from said second cassette. More preferably, the concatemer comprises cassettes, wherein substantially all cassettes are different. The difference between the cassettes may arise from differences between promoters, and/or expressible nucleotide sequences, and/or spacers, and/or terminators, and/or introns.

The number of cassettes in a single concatemer is largely determined by the host species into which the concatemer is eventually to be inserted and the vector through which the insertion is carried out. The concatemer thus may comprise at least 10 cassettes, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000 cassettes.

Each of the cassettes may be laid out as described above.

Once the concatemer has been assembled or concatenated it may be ligated into a suitable vector. Such a vector may advantageously comprise an artificial chromosome. The basic requirements for a functional artificial chromosome have been described in US 4,464,472, the contents of which is hereby incorporated by reference. An artificial chromosome or a functional minichromosome, as it may also be termed must comprise a DNA sequence capable of replication and stable mitotic maintenance in a host cell comprising a DNA segment coding for centromere-like activity during mitosis of said host and a DNA sequence coding for a replication site recognized by said host.

Suitable artificial chromosomes include a Yeast Artificial Chromosome (YAC) (see e.g. Murray et al, Nature 305:189-193; or US 4,464,472), a mega Yeast Artificial Chromosome (mega YAC), a Bacterial Artificial Chromosome (BAC), a mouse

artificial chromosome, a Mammalian Artificial Chromosome (MAC) (see e.g. US 6,133,503 or US 6,077,697), an Insect Artificial Chromosome (BUGAC), an Avian Artificial Chromosome (AVAC), a Bacteriophage Artificial Chromosome, a Baculovirus Artificial Chromosome, a plant artificial chromosome (US 5,270,201), a BIBAC vector (US 5,977,439) or a Human Artificial Chromosome (HAC).

The artificial chromosome is preferably so large that the host cell perceives it as a "real" chromosome and maintains it and transmits it as a chromosome. For yeast and other suitable host species, this will often correspond approximately to the size of the smallest native chromosome in the species. For Saccharomyces, the smallest chromosome has a size of 225 Kb.

MACs may be used to construct artificial chromosomes from other species, such as insect and fish species. The artificial chromosomes preferably are fully functional stable chromosomes. Two types of artificial chromosomes may be used. One type, referred to as SATACs [satellite artificial chromosomes] are stable heterochromatic chromosomes, and the other type are minichromosomes based on amplification of euchromatin.

Mammalian artificial chromosomes provide extra-genomic specific integration sites for introduction of genes encoding proteins of interest and permit megabase size DNA integration, such as integration of concatemers according to the invention.

According to another embodiment of the invention, the concatemer may be integrated into the host chromosomes or cloned into other types of vectors, such as a plasmid vector, a phage vector, a viral vector or a cosmid vector.

A preferable artificial chromosome vector is one that is capable of being conditionally amplified in the host cell, e.g. in yeast. The amplification preferably is at least a 10 fold amplification. Furthermore, it is advantageous that the cloning site of the artificial chromosome vector can be modified to comprise the same restriction site as the one bordering the cassettes described above, i.e. RS2 and/or RS2'.

Concatenation

Cassettes to be concatenated are normally excised from a vector either by digestion with restriction enzymes or by PCR. After excision the cassettes may be separated from the vector through size fractionation such as gel filtration or through tagging of known sequences in the cassettes. The isolated cassettes may then be joined together either through interaction between sticky ends or through ligation of blunt ends.

Single-stranded compatible ends may be created by digestion with restriction enzymes. For concatenation a preferred enzyme for excising the cassettes would be a rare cutter, i.e. an enzyme that recognises a sequence of 7 or more nucleotides. Examples of enzymes that cut very rarely are the meganucleases, many of which are intron encoded, like e.g. I-Ceu I, I-Sce I, I-Ppo I, and PI-Psp I (see eample 6d for more). Other preferred enzymes recognize a sequence of 8 nucleotides like e.g. Asc I, AsiS I, CciN I, CspB I, Fse I, MchA I, Not I, Pac I, Sbf I, Sda I, Sgf I, SgrA I, Sse232 I, and Sse8387 I, all of which create single stranded, palindromic compatible ends.

Other preferred rare cutters, which may also be used to control orientation of individual cassettes in the concatemer are enzymes that recognize non-palindromic sequences like e.g. Aar I, Sap I, Sfi I, Sdi I, and Vpa (see example 6c for more).

Alternatively, cassettes can be prepared by the addition of restriction sites to the ends, e.g. by PCR or ligation to linkers (short synthetic dsDNA molecules). Restriction enzymes are continuously being isolated and characterised and it is anticipated that many of such novel enzymes can be used to generate single-stranded compatible ends according to the present invention.

It is conceivable that single stranded compatible ends can be made by cleaving the vector with synthetic cutters. Thus, a reactive chemical group that will normally be able to cleave DNA unspecifically can cut at specific positions when coupled to another molecule that recognises and binds to specific sequences. Examples of molecules that recognise specific dsDNA sequences are DNA, PNA, LNA, phosphothioates, peptides, and amides. See e.g. Armitage, B.(1998) Chem. Rev.

98: 1171-1200, who describes photocleavage using e.g. anthraquinone and UV light; Dervan P.B. & Bürli R.W. (1999) Curr. Opin. Chem. Biol. 3: 688-93 describes the specific binding of polyamides to DNA; Nielsen, P.E. (2001) Curr. Opin. Biotechnol. 12: 16-20 describes the specific binding of PNA to DNA, and Chemical Reviews special thematic issue: RNA/DNA Cleavage (1998) vol. 98 (3) Bashkin J.K. (ed.) ACS publications, describes several examples of chemical DNA cleavers.

Single-stranded compatible ends may also be created by using e.g. PCR primers including dUTP and then treating the PCR product with Uracil-DNA glycosylase (Ref: US 5,035,996) to degrade part of the primer. Alternatively, compatible ends can be created by tailing both the vector and insert with complimentary nucleotides using Terminal Transferase (Chang, LMS, Bollum TJ (1971) J Biol Chem 246:909).

It is also conceivable that recombination can be used to generate concatemers, e.g. through the modification of techniques like the Creator™ system (Clontech) which uses the Cre-loxP mechanism (Sauer B 1993 Methods Enzymol 225:890-900) to directionally join DNA molecules by recombination or like the Gateway™ system (Life Technologies, US 5,888,732) using lambda *att* attachment sites for directional recombination (Landy A 1989, Ann Rev Biochem 58:913). It is envisaged that also lambda *cos* site dependent systems can be developed to allow concatenation.

More preferably the cassettes may be concatenated without an intervening purification step through excision from a vector with two restriction enzymes, one leaving sticky ends on the cassettes and the other one leaving blunt ends in the vectors. This is the preferred method for concatenation of cassettes from vectors having the basic structure of [RS1-RS2-SP-PR-X-TR-SP-RS2'-RS1'].

An alternative way of producing concatemers free of vector sequences would be to PCR amplify the cassettes from a single stranded primary vector. The PCR product must include the restriction sites RS2 and RS2' which are subsequently cleaved by its cognate enzyme(s). Concatenation can then be performed using the digested PCR product, essentially without interference from the single stranded primary vector template or the small double stranded fragments, which have been cut from the ends.

The concatemer may be assembled or concatenated by concatenation of at least two cassettes of nucleotide sequences each cassette comprising a first sticky end, a spacer sequence, a promoter, an expressible nucleotide sequence, a terminator, a spacer sequence, and a second sticky end. A flow chart of the procedure is shown in figure 2a.

Preferably concatenation further comprises

starting from a primary vector [RS1-RS2-SP-PR-X-TR-SP-RS2'-RS1'],
wherein X denotes an expressible nucleotide sequence,
RS1 and RS1' denote restriction sites,
RS2 and RS2' denote restriction sites different from RS1 and RS1',
SP individually denotes a spacer sequence of at least two nucleotides,
PR denotes a promoter,
TR denotes a terminator,

i) cutting the primary vector with the aid of at least one restriction enzyme specific for RS2 and RS2' obtaining cassettes having the general formula [rs_2-SP-PR-X-TR-SP-rs_1] wherein rs_1 and rs_2 together denote a functional restriction site RS2 or RS2',

ii) assembling the cut out cassettes through interaction between rs_1 and rs_2.

In this way at least 10 cassettes can be concatenated, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000.

According to an especially preferred embodiment, vector arms each having a RS2 or RS2' in one end and a non-complementary overhang or a blunt end in the other end are added to the concatenation mixture together with the cassettes described above to further simplify the procedure (see Fig. 2b). One example of a suitable vector for providing vector arms is disclosed in Fig. 7 TRP1, URA3, and HIS3 are auxotrophic marker genes, and AmpR is an E. coli antibiotic marker gene. CEN4 is a centromer and TEL are telomeres. ARS1 and PMB1 allow replication in yeast and E. coli respectively. BamH I and Asc I are restriction enzyme recognition sites. The

nucleotide sequence of the vector is set forth in SEQ ID NO 4. The vector is digested with BamHI and AscI to liberate the vector arms, which are used for ligation to the concatemer.

The ratio of vector arms to cassettes determines the maximum number of cassettes in the concatemer as illustrated in figure 8. The vector arms preferably are artificial chromosome vector arms such as those described in Fig. 7.

It is of course also possible to add stopper fragments to the concatenation solution, the stopper fragments each having a RS2 or RS2' in one end and a non-complementary overhang or a blunt end in the other end. The ratio of stopper fragments to cassettes can likewise control the maximum size of the concatemer.

The complete sequence of steps to be taken when starting with the isolation of mRNA until inserting into an entry vector may include the following steps

i) isolating mRNA from an expression state,

ii) obtaining substantially full length cDNA corresponding to the mRNA sequences,

iii) inserting the substantially full length cDNA into a cloning site in a cassette in a primary vector, said cassette being of the general formula in 5'→3' direction:
[RS1-RS2-SP-PR-CS-TR-SP-RS2'-RS1']
wherein CS denotes a cloning site.

In preparation of the concatemer, genes may be isolated from different entry libraries to provide the desired selection of genes. Accordingly, concatenation may further comprise selection of vectors having expressible nucleotide sequences from at least two different expression states, such as from two different species. The two different species may be from two different classes, such as from two different divisions, more preferably from two different sub-kingdoms, such as from two different kingdoms.

As an alternative to including vector arms in the concatenation reaction it is possible to ligate the concatemer into an artificial chromosome selected from the group

comprising yeast artificial chromosome, mega yeast artificial chromosome, bacterial artificial chromosome, mouse artificial chromosome, human artificial chromosome.

Preferably at least one inserted concatemer further comprises a selectable marker. The marker(s) are conveniently not included in the concatemer as such but rather in an artificial chromosome vector, into which the concatemer is inserted. Selectable markers generally provide a means to select, for growth, only those cells which contain a vector. Such markers are of two types: drug resistance and auxotrophy. A drug resistance marker enables cells to grow in the presence of an otherwise toxic compound. Auxotrophic markers allow cells to grow in media lacking an essential component by enabling cells to synthesise the essential component (usually an amino acid).

Illustrative and non-limiting examples of common compounds for which selectable markers are available with a brief description of their mode of action follow:

Prokaryotic

- Ampicillin: interferes with a terminal reaction in bacterial cell wall synthesis. The resistance gene (bla) encodes beta-lactamase which cleaves the beta-lactam ring of the antibiotic thus detoxifying it.
- Tetracycline: prevents bacterial protein synthesis by binding to the 30S ribosomal subunit. The resistance gene (tet) specifies a protein that modifies the bacterial membrane and prevents accumulation of the antibiotic in the cell.
- Kanamycin: binds to the 70S ribosomes and causes misreading of messenger RNA. The resistant gene (nptH) modifies the antibiotic and prevents interaction with the ribosome.
- Streptomycin: binds to the 30S ribosomal subunit, causing misreading of messenger RNA. The resistance gene (Sm) modifies the antibiotic and prevents interaction with the ribosome.
- Zeocin: this new bleomycin-family antibiotic intercalates into the DNA and cleaves it. The Zeocin resistance gene encodes a 13,665 dalton protein. This protein confers resistance to Zeocin by binding to the antibiotic and preventing it from binding DNA. Zeocin is effective on most aerobic cells and can be used for selection in mammalian cell lines, yeast, and bacteria.

Eukaryotic

- Hygromycin: a aminocyclitol that inhibits protein synthesis by disrupting ribosome translocation and promoting mistranslation. The resistance gene (hph) detoxifies hygromycin -B- phosphorylation.
- Histidinol: cytotoxic to mammalian cells by inhibiting histidyl-tRNA synthesis in histidine free media. The resistance gene (hisD) product inactivates histidinol toxicity by converting it to the essential amino acid, histidine.
- Neomycin (G418): blocks protein synthesis by interfering with ribosomal functions. The resistance gene ADH encodes amino glycoside phosphotransferase which detoxifies G418.
- Uracil: Laboratory yeast strains carrying a mutated gene which encodes orotidine -5'- phosphate decarboxylase, an enzyme essential for uracil biosynthesis, are unable to grow in the absence of exogenous uracil. A copy of the wild-type gene (ura4+, S. pombe or URA3 S. cerevisiae) carried on the vector will complement this defect in transformed cells.
- Adenosine: Laboratory strains carrying a deficiency in adenosine synthesis may be complemented by a vector carrying the wild type gene, ADE 2.
- Amino acids: Vectors carrying the wild-type genes for LEU2, TRP 1, HIS 3 or LYS 2 may be used to complement strains of yeast deficient in these genes.
- Zeocin: this new bleomycin-family antibiotic intercalates into the DNA and cleaves it. The Zeocin resistance gene encodes a 13,665 dalton protein. This protein confers resistance to Zeocin by binding to the antibiotic and preventing it from binding DNA. Zeocin is effective on most aerobic cells and can be used for selection in mammalian cell lines, yeast, and bacteria.

Transgenic cells

In one aspect of the invention, the concatemers comprising the multitude of cassettes are introduced into a host cell, in which the concatemers can be maintained and the expressible nucleotide sequences can be expressed in a co-ordinated way. The cassettes comprised in the concatemers may be isolated from the host cell and re-assembled due to their uniform structure with –preferably – concatemer restriction sites between the cassettes.

The host cells selected for this purpose are preferably cultivable under standard laboratory conditions using standard culture conditions, such as standard media and protocols. Preferably the host cells comprise a substantially stable cell line, in which the concatemers can be maintained for generations of cell division. Standard techniques for transformation of the host cells and in particular methods for insertion of artificial chromosomes into the host cells are known.

It is also of advantage if the host cells are capable of undergoing meiosis to perform sexual recombination. It is also advantageous that meiosis is controllable through external manipulations of the cell culture. One especially advantageous host cell type is one where the cells can be manipulated through external manipulations into different mating types.

The genome of a number of species have already been sequenced more or less completely and the sequences can be found in databases. The list of species for which the whole genome has been sequenced increases constantly. Preferably the host cell is selected from the group of species, for which the whole genome or essentially the whole genome has been sequenced. The host cell should preferably be selected from a species that is well described in the literature with respect to genetics, metabolism, physiology such as model organism used for genomics research.

The host organism should preferably be conditionally deficient in the abilities to undergo homologous recombination. The host organism should preferably have a codon usage similar to that of the donor organisms. Furthermore, in the case of genomic DNA, if eukaryotic donor organisms are used, it is preferable that the host organism has the ability to process the donor messenger RNA properly, e.g., splice out introns.

The host cells can be bacterial, archaebacteria, or eukaryotic and can constitute a homogeneous cell line or mixed culture. Suitable cells include the bacterial and eukaryotic cell lines commonly used in genetic engineering and protein expression.

Preferred prokaryotic host organisms may include but are not limited to Escherichia coli, Bacillus subtilis, B licehniformis, B. cereus, Streptomyces lividans,

Streptomyces coelicolor, Pseudomonas aeruginosa, Myxococcus xanthus. Rhodococcus, Streptomycetes, Actinomycetes, Corynebacteria, Bacillus, Pseudomonas, Salmonella, and Erwinia. The complete genome sequences of E. coli and Bacillus subtilis are described by Blattner et al., Science 277, 1454-1462 (1997); Kunst et al., Nature 390, 249-256 (1997)).

Preferred eukaryotic host organisms are mammals, fish, insects, plants, algae and fungi.

Examples of mammalian cells include those from, e.g., monkey, mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Preferred mammalian host cells include but are not limited to those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa etc (see Kriegler M. in "Gene Transfer and Expression: A Laboratory Manual", New York, Freeman & Co. 1990), and stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, kidney, liver, muscle, and skin cells.

Examples of insect cells include baculo lepidoptera.

Examples of plant cells include maize, rice, wheat, cotton, soybean, and sugarcane. Plant cells such as those derived from Nicotiana and Arabidopsis are preferred

Examples of fungi include penicillium, aspergillus, such as Aspergillus nidulans, podospora, neurospora, such as Neurospora crassa, saccharomyces, such as Saccharomyces cerevisiae (budding yeast), Schizosaccharomyces, such as Schizosaccharomyces pombe (fission yeast), Pichia spp, such as Pichia pastoris, and Hansenula polymorpha (methylotropic yeasts).

In a preferred embodiment the host cell is a yeast cell, and an illustrative and not limiting list of suitable yeast host cells comprise: baker's yeast, Kluyveromyces marxianus, K. lactis, Candida utilis, Phaffia rhodozyma, Saccharomyces boulardii, Pichia pastoris, Hansenula polymorpha, Yarrowia lipolytica, Candida paraffinica, Schwanniomyces castellii, Pichia stipitis, Candida shehatae, Rhodotorula glutinis, Lipomyces lipofer, Cryptococcos curvatus, Candida spp. (e.g. C. palmioleophila),

Yarrowia lipolytica, Candida guilliermondii, Candida, Rhodotorula spp., Saccharomycopsis spp., Aureobasidium pullulans, Candida brumptii, Candida hydrocarbofumarica, Torulopsis, Candida tropicalis, Saccharomyces cerevisiae, Rhodotorula rubra, Candida flaveri, Eremothecium ashbyii, Pichia spp., Pichia pastoris, Kluyveromyces, Hansenula, Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola.

The choice of host will depend on a number of factors, depending on the intended use of the engineered host, including pathogenicity, substrate range, environmental hardiness, presence of key intermediates, ease of genetic manipulation, and likelihood of promiscuous transfer of genetic information to other organisms. Particularly advantageous hosts are E. coli, lactobacilli, Streptomycetes, Actinomycetes, Saccharomyces and filamentous fungi.

In any one host cell it is possible to make all sorts of combinations of expressible nucleotide sequences from all possible sources. Furthermore, it is possible to make combinations of promoters and/or spacers and/or introns and/or terminators in combination with one and the same expressible nucleotide sequence.

Thus in any one cell there may be expressible nucleotide sequences from two different expression states. Furthermore, these two different expression states may be from one species or advantageously from two different species. Any one host cell may also comprise expressible nucleotide sequences from at least three species, such as from at least four, five, six, seven, eight, nine or ten species, or from more than 15 species such as from more than 20 species, for example from more than 30, 40 or 50 species, such as from more than 100 different species, for example from more than 300 different species, such as form more than 500 different species, for example from more than 1000 different species, thereby obtaining combinations of large numbers of expressible nucleotide sequences from a large number of species. In this way potentially unlimited numbers of combinations of expressible nucleotide sequences can be combined across different expression states. These different expression states may represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera. The different species may be from at least two different phylae, such as from at least two different

classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.

Any two of these species may be from two different classes, such as from two different divisions, more preferably from two different sub-kingdoms, such as from two different kingdoms. Thus expressible nucleotide sequences may be combined from a eukaryot and a prokaryot into one and the same cell.

According to another embodiment of the invention, the expressible nucleotide sequences may be from one and the same expression state. The products of these sequences may interact with the products of the genes in the host cell and form new enzyme combinations leading to novel biochemical pathways. Furthermore, by putting the expressible nucleotide sequences under the control of a number of promoters it becomes possible to switch on and off groups of genes in a co-ordinated manner. By doing this with expressible nucleotide sequences from only one expression states, novel combinations of genes are also expressed.

The number of concatemers in one single cell may be at least one concatemer per cell, preferably at least 2 concatemers per cell, more preferably 3 per cell, such as 4 per cell, more preferably 5 per cell, such as at least 5 per cell, for example at least 6 per cell, such as 7, 8, 9 or 10 per cell, for example more than 10 per cell. As described above, each concatemer may preferably comprise up to 1000 cassettes, and it is envisages that one concatemer may comprise up to 2000 cassettes. By inserting up to 10 concatemers into one single cell, this cell may thus be enriched with up to 20,000 heterologous expressible genes, which under suitable conditions may be turned on and off by regulation of the regulatable promoters.

Often it is more preferable to provide cells having anywhere between 10 and 1000 heterologous genes, such as 20-900 heterologous genes, for example 30 to 800 heterologous genes, such as 40 to 700 heterologous genes, for example 50 to 600 heterologous genes, such as from 60 to 300 heterologous genes or from 100 to 400 heterologous genes which are inserted as 2 to 4 artificial chromosomes each containing one concatemer of genes. The genes may advantageously be located on 1 to 10 such as from 2 to 5 different concatemers in the cells. Each concatemer may advantageously comprise from 10 to 1000 genes, such as from 10 to 750 genes,

such as from 10 to 500 genes, such as from 10 to 200 genes, such as from 20 to 100 genes, for example from 30 to 60 genes, or from 50 to 100 genes.

The concatemers may be inserted into the host cells according to any known transformation technique, preferably according to such transformation techniques that ensure stable and not transient transformation of the host cell. The concatemers may thus be inserted as an artificial chromosome which is replicated by the cells as they divide or they may be inserted into the chromosomes of the host cell. The concatemer may also be inserted in the form of a plasmid such as a plasmid vector, a phage vector, a viral vector, a cosmid vector, that is replicated by the cells as they divide. Any combination of the three insertion methods is also possible. One or more concatemers may thus be integrated into the chromosome(s) of the host cell and one or more concatemers may be inserted as plasmids or artificial chromosomes. One or more concatemers may be inserted as artificial chromosomes and one or more may be inserted into the same cell via a plasmid.

Examples

Example 1

In the examples 1-3 an Asc1 site was introduced into the EcoR1 site in pYAC4 (Sigma, Burke DT et al. 1987, Science vol 236, p 806), so that sticky ends match the Asc1 site(= RS2 in general formula of this patent) of the cassettes in pEVE vectors

Preparation of EVACs (EVolvable Artificial Chromosomes) including size fractioning

preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB (sigma) with a single colony of E. coli DH5α containing pYAC4-Asc
2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
3. digest 100µg pYAC4-Asc w. BamH1 and Asc1
4. dephosphorylate fragments and heat inactivate phosphatase(20 min, 80 C)
5. purify fragments(e.g. Qiaquick Gel Extraction Kit)
6. run 1 % agarose gel to estimate amount of fragment

Preparation of expression cassettes

1. take 100 µg of plasmid preparation from each of the following libraries
   a) pMA-CAR
   b) pCA-CAR
   c) Phaffia cDNA library
   d) Carrot cDNA library
2. digest w. Srf1(10 units/prep, 37C overnight)
3. dephosphorylate (10 units/prep, 37C, 2h)
4. heat inactivate 80C, 20 min
5. concentrate and change buffer (precipitation or ultra filtration),
6. digest w. Asc1. (10 units/prep, 37 C, overnight)
7. adjust volume of preps to 100 µL

preparation of EVACs

Different types of EVACs have been made by varying the ratio of the different libraries that goes into the ligation reaction.

	pMA-CAR	pCA-CAR	Phaffia cDNA	Carrot cDNA
EVAC				
A	40%	40%	10%	10%
B	25%	25%	25%	25%

1. add ~100 ng arms of pYAC4-Asc /100 µg of cassette mixture
2. concentrate to < 33.5 µL
3. add 2.5 units of T4 DNA-ligase + 4 µL 10x ligase buffer. Adjust to 40 µL
4. ligate 3 h, 16 C
5. stop reaction by adding 2 µL of 500 mM EDTA
6. bring reaction volume to 125 µL, add 25 µL loading mix, heat at 60C for 5 min
7. distribute evenly in 10 wells of a 1% LMP agarose gel
8. run pulsed field gel (CHEF III, 1% LMP agarose, ½ strength TBE (BioRad), angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s, run time 30 h)

9. stain part of the gel that contains molecular weight markers + 1 sample lane for quality check
10. cut remaining 9 sample lanes corresponding to mw. 97 – 194 kb(fraction 1); 194 – 291 kb(fraction 2); 291-365 kb(fraction 3) from the gel
11. agarase gel in high NaCl agarase buffer . 1 u agarase / 100µg gel. 40C 3 h
12. concentrate preparation to < 20 µL
13. transform suitable yeast strain w. preparation using alkali/cation transformation
14. plate on selective minimal media plates
15. incubate 30 C for 4-5 days
16. pick colonies
17. analyse colonies

Example 2

Preparation of EVACs (EVolvable Artificial Chromosomes) with direct transformation

preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB with a single colony of DH5α containing pYAC4-Asc
2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
3. digest 100µg pYAC4-Asc w. BamH1 and Asc1
4. dephosphorylate fragments and heat inactivate phosphatase(20 min, 80 C)
5. purify fragments(e.g. Qiaquick Gel Extraction Kit)
1. run 1 % agarose gel to estimate amount of fragment

Preparation of expression cassettes

1. take 100 µg of plasmid preparation from each of the following libraries
   e) pMA-CAR
   f) pCA-CAR
   g) Phaffia cDNA library
   h) Carrot cDNA library
2. digest w. Srf1(10 units/prep, 37C overnight)
3. dephosphorylate (10 units/prep, 37C, 2h)
4. heat inactivate 80C, 20 min
5. concentrate and change buffer (precipitation or ultra filtration),
6. digest w. Asc1. (10 units/prep, 37 C, overnight)

7. adjust volume of preps to 100 µL

preparation of EVACs

Different types of EVACs have been made by varying the ratio of the different libraries that goes into the ligation reaction.

	pMA-CAR	pCA-CAR	Phaffia cDNA	Carrot cDNA
EVAC				
A	40%	40%	10%	10%
B	25%	25%	25%	25%

1. concentrate to < 32 µL
2. add 1 unit of T4 DNA-ligase + 4 µL 10x ligase buffer. Adjust to 40 µL
3. ligate 2 h, 16 C
4. stop reaction by adding 2 µL of 500 mM EDTA, heat inactivate 60C, 20 min
5. bring reaction volume to 500 µL with dH_2O, concentrate to 30 µL
6. add 10 U Asc1, 4 µL 10X Asc1 buffer, bring to 40 µL
7. incubate at 37C for 1h (alternatively 15 min 30 min)
8. heat inactivate 60C, 20 min
9. add 2 µg YAC4-Asc arms, 1 U T4 DNA ligase, 10 µL 10X ligase buffer, bring to 100 µL
10. incubate ON, 16C
11. add water to 500 µL
12. concentrate to 25 µL
13. transform suitable yeast strain w. preparation using alkali/cation transformation or other suitable transformation method
14. plate on selective minimal media plates
15. incubate 30 C for 4-5 days
16. pick colonies
17. analyse colonies

Example 3

Preparation of EVACs (EVolvable Artificial Chromosomes) (Small scale preparation)

Preparation of expression cassettes

1. inoculate 5 ml of LB-medium (Sigma) with library inoculum corresponding to a 10+ fold representation of library. Grow overnight
2. make plasmid miniprep from 1.5 ml of culture (E.g. Qiaprep spin miniprep kit)
3. digest plasmid w. Srf 1
4. dephosphorylate fragments and heat inactivate phosphatase(20 min, 80 C)
5. digest w. Asc1
6. run 1/10 of reaction in 1% agarose to estimate amount of fragment

preparation of pYAC4-Asc arms

1. inoculate 150 ml of LB with a single colony of E. coli DH5α containing pYAC4-Asc
2. grow to OD600 ~ 1, harvest cells and make plasmid preparation
3. digest 100µg pYAC4-Asc w. BamH1 and Asc1
4. dephosphorylate fragments and heat inactivate phosphatase(20 min, 80 C)
5. purify fragments(E.g. Qiaquick Gel Extraction Kit)
6. run 1 % agarose gel to estimate amount of fragment

preparation of EVACs

1. mix expression cassette fragments with YAC-arms so that cassette/arm ration is ~1000/1
2. if needed concentrate mixture(use e.g. Microcon YM30) so fragment concentration > 75 ng/µL reaction
3. add 1 U T4 DNA ligase, incubate 16C, 1-3 h . Stop reaction by adding 1 µL of 500 mM EDTA
4. run pulsed field gel (CHEF III, 1% LMP agarose, ½ strength TBE, angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s, run time 30 h) Load sample in 2 lanes.
5. stain part of the gel that contains molecular weight markers
6. cut sample lanes corresponding to mw. 100 – 200 kb
7. agarase gel in high NaCl agarase buffer . 1 U agarase / 100 mg gel

8. concentrate preparation to < 20 µL
9. transform suitable yeast strain w. preparation using electroporation
10. plate on selective minimal media plates
11. incubate 30 C for 4-5 days
12. pick colonies

Example 4: cDNA libraries used in the production of EVACs

1. *Daucus carota*, carrot root library:
   - Full length
   - Oligo dT primed, directional cDNA library
   - cDNA library made using a pool of 3 Evolva EVE 4, 5 & 8 vectors (Fig. 4, 5, 6)
   - Number of independent clones: 41.6×10^6
   - Average size: 0.9 – 2.9 kb
   - Number of different genes present: 5000 -10000

2. *Xanthophyllomyces dendrorhous*, (yeast), hole organism library
   - Full length
   - Oligo dT primed, directional cDNA library
   - cDNA library made using a pool of 3 Evolva EVE 4, 5 & 8 vectors (Fig. 4, 5, 6)
   - Number of independent clones: 48.0×10^6
   - Average size: 1.0 – 3.8 kb
   - Number of different genes present: 5000 -10000

3. Target carotenoid gene cDNA library
   - Full length and normalised
   - Directional cDNA cloning
   - Library made by cloning each gene individually in 2 Evolva EVE 4, 5 & 8 vectors (Fig. 4, 5, 6)
   - Number of different genes: 48
   - Species and genes used:
     - Gentiana sp., ggps, psy, pds, zds, lcy-b, lcy-e, bhy, zep
     - Rhodobacter capsulatus, idi, crtC, crtF
     - Erwinia uredovora, crtE, crtB, crtI, crtY, crtZ

- Nostoc anabaena, zds
- Synechococcus PCC7942, pds
- Erwinia herbicola, crtE, crtB, crtI, crtY, crtZ
- Staphylococcus aureus, crtM, crtN
- Xanthophyllomyces dendrorhous, crtI, crtYb
- Capsicum annuum, ccs, crtL
- Nicotiana tabacum, crtL, bchy
- Prochlorococcus sp., lcy-b, lcy-e
- Saccharomyces cerevisiae, idi
- Corynebacterium sp., crtI, crtYe, crtYf, crtEb
- Lycopersicon esculentum, psy-1
- Neurospora crassa, al1

Example 5: Transformation of EVACs

Example 5a: Transformation

1. Inoculate a single colony into 100 ml YPD broth and grow with aeration at 30°C to mid log, 2 x 10^6 to 2 x 10^7 cells/ml.
2. Spin to pellet cells at 400 x g for 5 minutes; discard supernatant.
3. Resuspend cells in a total of 9 ml TE, pH 7.5. Spin to pellet cells and discard supernatant.
4. Gently resuspend cells in 5 ml 0.1 M Lithium/Cesium Acetate solution, pH 7.5.
5. Incubate at 30°C for 1 hour with gentle shaking.
6. Spin at 400 x g for 5 minutes to pellet cells and discard supernatant.
7. Gently resuspend in 1 ml TE, pH 7.5. Cells are now ready for transformation.
8. In a 1.5 ml tube combine:
   - 100 µl yeast cells
   - 5 µl Carrier DNA (10 mg/ml)
   - 5 µl Histamine Solution
   - 1/5 of an EVAC preparation in a 10 µl volume (max). (One EVAC preparation is made of 100 µg of concatenation reaction mixture)
9. Gently mix and incubate at room temperature for 30 minutes.
10. In a separate tube, combine 0.8 ml 50% (w/v) PEG 4000 and 0.1 ml TE and 0.1 ml of 1 M LiAc for each transformation reaction. Add 1 ml of this PEG/TE/LiAc mix to each transformation reaction. Mix cells into solution with gentle pipetting.

11. Incubate at 30°C for 1 hour.
12. Heat shock at 42°C for 15 minutes; cool to 30°C.
13. Pellet cells in a microcentrifuge at high speed for 5 seconds and remove supernatant.
14. Resuspend in 200 µl of rich media and plate in appropriate selective media
15. Incubate at 30°C for 48-72 hours until transformant colonies appear.

Example 5b: Transformation of EVACs using electroporation

100 ml of YPD is inoculated with one yeast colony and grown to OD_{600} = 1.3 to 1.5. The culture is harvested by centrifuging at 4000 × g and 4°C. The cells are resuspended in 16 ml sterile H_2O. Add 2 ml 10 × TE buffer, pH 7.5 and swirl to mix. Add 2 ml 10 × lithium acetate solution (1 M, pH 7.5) and swirl to mix. Shake gently 45 min at 30°C. Add 1.0 ml 0.5 M DTE while swirling. Shake gently 15 min at 30°C. The yeast suspension is diluted to 100 ml with sterile water. The cells are washed and concentrated by centrifuging at 4000 × g, resuspending the pellet in 50 ml ice-cold sterile water, centrifuging at 4000 × g, resuspending the pellet in 5 ml ice-cold sterile water, centrifuging at 4000 × g and resuspending the pellet in 0.1 ml ice-cold sterile 1 M sorbitol. The electroporation was done using a *Bio-Rad Gene Pulser*. In a sterile 1.5-ml microcentrifuge tube 40 µl concentrated yeast cells were mixed with 5 µl 1:10 diluted EVAC preparation. The yeast-DNA mix is transferred to an ice-cold 0.2-cm-gap disposable electroporation cuvette and pulsed at 1.5 kV, 25 µF, 200 Ω. 1 ml ice-cold 1 M sorbitol is added to the cuvette to recover the yeast. Aliquots are spread on selective plates containing 1 M sorbitol. Incubate at 30°C until colonies appear.

Example 6: Rare restriction enzymes with recognition sequence and cleavage points

In this example, rare restriction enzymes are listed together with their recognition sequence and cleavage points. (^) indicates cleavage points 5'-3' sequence and (_) Indicates cleavage points in the complementary sequence.

W = A or T; N = A, C, G, or T

6a) Unique, palindromic overhang

AscI	GG^CGCG_CC
AsiSI	GCG_AT^CGC
CciNI	GC^GGCC_GC
CspBI	GC^GGCC_GC
FseI	GG_CCGG^CC
MchAI	GC^GGCC_GC
NotI	GC^GGCC_GC
PacI	TTA_AT^TAA
SbfI	CC_TGCA^GG
SdaI	CC_TGCA^GG
SgfI	GCG_AT^CGC
SgrAI	CR^CCGG_YG
Sse232I	CG^CCGG_CG
Sse8387I	CC_TGCA^GG

6b) No overhang

BstRZ246I	ATTT^AAAT
BstSWI	ATTT^AAAT
MspSWI	ATTT^AAAT
MssI	GTTT^AAAC
PmeI	GTTT^AAAC
SmiI	ATTT^AAAT
SrfI	GCCC^GGGC
SwaI	ATTT^AAAT

6c) Non-palindromic and/or variable overhang

AarI	CACCTGCNNNN^NNNN_
AbeI	CC^TCA_GC
AloI	^NNNNN_NNNNNNNGAACNNNNNNTCCNNNNNNN_NNNNN^
BaeI	^NNNNN_NNNNNNNNNNACNNNNGTAYCNNNNNNN_NNNNN^
BbvCI	CC^TCA_GC
CpoI	CG^GWC_CG
CspI	CG^GWC_CG
Pfl27I	RG^GWC_CY
PpiI	^NNNNN_NNNNNNNGAACNNNNNCTCNNNNNNNN_NNNNN^
PpuMI	RG^GWC_CY
PpuXI	RG^GWC_CY
Psp5II	RG^GWC_CY
PspPPI	RG^GWC_CY
RsrII	CG^GWC_CG
Rsr2I	CG^GWC_CG
SanDI	GG^GWC_CC
SapI	GCTCTTCN^NNN_
SdiI	GGCCN_NNN^NGGCC
SexAI	A^CCWGG_T
SfiI	GGCCN_NNN^NGGCC
Sse1825I	GG^GWC_CC
Sse8647I	AG^GWC_CT
VpaK32I	GCTCTTCN^NNN_

6d) Meganucleases

I-Sce I	TAGGGATAA_CAGG^GTAAT
I-Ceu I	ACGGTC_CTAA^GGTAG

I-Cre I	AAACGTC_GTGA^GACAGTTT
I-Sce II	GGTC_ACCC^TGAAGTA
I-Sce III	GTTTTGG_TAAC^TATTTAT
Endo. Sce I	GATGCTGC_AGGC^ATAGGCTTGTTTA
PI-Sce I	GG_GTGC^GGAGAA
PI-Psp I	TGGCAAACAGCTA_TTAT^GGGTATTATGGGT
I-Ppo I	CTCTC_TTAA^GGTAG
HO	TTTCCGC_AACA^GT
I-Tev I	NN_NN^NNTCAGTAGATGTTTTTCTTGGTCTACCGTTT

More meganucleases have been identified, but their precise sequence of recognition has not been determined, see e.g. www.meganuclease.com

Example 7: Concatemer size limitation experiments (use of stoppers)

Materials used:

pYAC4 (Sigma. Burke et al. 1987, science, vol 236, p.806) was digested w. EcoR1 and BamH1 and dephosphorylated

pSE420 (invitrogen) was linearised using EcoR1 and used as the model fragment for concatenation.

T4 DNA ligase (Amersham-pharmacia biotech) was used for ligation according to manufacturers instructions.

Method: Fragments and arms were mixed in the ratios(concentrations are arbitrary units) indicated on figures 9a and 9b. Ligation was allowed to proceed for 1 h at 16C. Reaction was stopped by the addition of 1 µL 500 mM EDTA. Products were analysed by standard agarose GE (1 % agarose, ½ strength TBE) or by PFGE(CHEF III, 1% LMP agarose, ½ strength TBE, angle 120, temperature 12 C, voltage 5.6V/cm, switch time ramping 5 – 25 s, run time 30 h)

The results are shown in Figure 9, wherein it is shown that the size of concatemers is proportional to the ratio of cassettes per YAC arms.

Example 8: Integration of expression cassettes into artificial chromosomes

Integration of expression cassettes into YAC12 was done essentially as done by Sears D.D., Hieter P., Simchen G., Genetics, 1994, 138, 1055-1065.

An AscI site was introduced into the Bgl II site of the integration vectors pGS534 and pGS525.

A β-galactosidase gene, as well as crtE, crtB, crtI and crtY from *Erwinia Uredovora* were cloned into pEVE4. These expression cassettes were ligated into AscI of the modified integration vectors pGS534 and pGS525.

Linearised pGS534 and pGS525 containing the expression cassettes were transformed into haploid yeast strains containing the appropriate target YAC which carries the Ade" gene. Red Ade- transformants were selected (the parent host strain is red due to the ade2-101 mutation).

Additional confirmation of correct integration of the β-galactosidase expression cassette was done using a β-galactosidase assay.

Example 9: Re-transformation of cells that already contain Artificial chromosomes to obtain at least 2 artificial chromosomes per cell

Yeast strains containing YAC12, Sears D.D., Hieter P., Simchen G., Genetics, 1994, 138, 1055-1065 were transformed with EVACs following the protocol described in example 4a. The transformed cells were plated on plates that select for cells that contained both YAC12 and EVACs.

Example 10: Example of different expression patterns "phenotypes" obtained using the same yeast clones under different expression conditions:

Colonies were picked with a sterile toothpick and streaked sequentially onto plates corresponding to the four repressed and/or induced conditions (-Ura/-Trp, -Ura/-Trp/-Met, -Ura/-Trp/+200 µM Cu_2SO_4, -Ura/-Trp/-Met/+200 µM Cu_2SO_4). 20 mg adenin was added to the media to suppress the ochre phenotype.

Claims

1. An artificial chromosome comprising at least one nucleotide concatemer, the concatemer comprising in the 5'→3' direction cassettes of nucleotide sequences of the general formula

$$[rs_2\text{-SP-PR-X-TR-SP-}rs_1]_n$$

wherein

rs_1 and rs_2 together denote a restriction site,
SP denotes a spacer of at least two nucleotide bases,
PR denotes a promoter, capable of functioning in a cell,
X denotes an expressible nucleotide sequence,
TR denotes a terminator, and
SP denotes a spacer of at least two nucleotide bases, and
$n \geq 2$.

2. The artificial chromosome according to claim 1, wherein the nucleotide sequence comprises a DNA sequence selected from the group comprising cDNA, genomic DNA.

3. The artificial chromosome according to claim 1, wherein the nucleotide sequence is single stranded, or partly single stranded.

4. The artificial chromosome according to claim 1, wherein the nucleotide sequence is double stranded.

5. The artificial chromosome according to any of the preceding claims, comprising nucleotide sequences from one expression state.

6. The artificial chromosome according to any of the preceding claims 1 to 4, comprising nucleotide sequences from at least two expression states.

7. The artificial chromosome according to any of the preceding claims, wherein the rs1-rs2 restriction site of at least two cassettes are recognised by the same restriction enzyme, more preferably are identical.

8. The artificial chromosome according to claim 7, wherein the rs1-rs2 restriction site of essentially all cassettes are recognised by the same restriction enzyme, more preferably are identical.

9. The artificial chromosome according to any of the preceding claims, wherein substantially all cassettes are different.

10. The artificial chromosome according to any of the preceding claims, wherein the difference comprises different promoters, and/or different expressible nucleotide sequences, and/or different spacers and/or different terminators and/or different introns.

11. The artificial chromosome according to any of the preceding claims, wherein n is at least 10, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000.

12. The artificial chromosome according to any of the preceding claims, wherein the artificial chromosome is selected from the group comprising a Yeast Artificial Chromosome, a mega Yeast Artificial Chromosome, a Bacterial Artificial Chromosome, a mouse artificial chromosome, a Mammalian Artificial Chromosome, an Insect Artificial Chromosome, an Avian Artificial Chromosome, a Bacteriophage Artificial Chromosome, a Baculovirus Artificial Chromosome, or a Human Artificial Chromosome.

13. The artificial chromosome according to any of the preceding claims, wherein the chromosome further comprises at least one selectable genetic marker, such as a recessive or a dominant marker.

14. The artificial chromosome according to claim 13, comprising at least two selectable genetic markers.

15. The artificial chromosome according to claim 13 to 14, wherein the at least one marker comprises a marker selected from the group comprising LEU 2, TRP 1, HIS 3, LYS 2, URA 3, ADE 2, Amyloglucosidase, β-lactamase, CUP 1, $G418^R$, TUN^R, KILk1, C230, SMR1, SFA, $Hygromycin^R$, $methotrexate^R$, $chloramphenicol^R$, $Diuron^R$, $Zeocin^R$, $Canavanine^R$.

16. The artificial chromosome according to any of the preceding claims, being designed to minimise the level of repeat sequences occurring in the concatemer.

17. The artificial chromosome according to any of the preceding claims, further comprising an intron sequence between the promoter and the expressible nucleotide sequence.

18. The artificial chromosome according to any of the preceding claims, wherein the restriction site comprises a restriction site from the list Example 6.

19. The artificial chromosome according to claim 18, wherein the restriction site comprises at least 6 bases such as at least 8 bases, for example at least 10 bases.

20. The artificial chromosome according to any of the preceding claims, wherein the GC content of the restriction site is more than 40%, preferably more than 50%, more preferably equal to or more than 60%.

21. The artificial chromosome according to any of the preceding claims, wherein the restriction enzyme recognising the restriction site produces sticky ends upon cleavage of a double stranded nucleotide sequence, preferably wherein the sticky ends have a pre-determined nucleotide sequence.

22. The artificial chromosome according to any of the preceding claims, further comprising a spacer sequence between TR and rs_2.

23. The artificial chromosome according to any of the preceding claims, wherein the spacer and the optional spacer sequence together comprise at least 50 bases, such as at least 60 bases, for example at least 75 bases, such as at least 100 bases, for example at least 150 bases, such as at least 200 bases, for example at least 250 bases, such as at least 300 bases, for example at least 400 bases, for example at least 500 bases, such as at least 750 bases, for example at least 1000 bases, such as at least 1100 bases, for example at least 1200 bases, such as at least 1300 bases, for example at least 1400 bases, such as at least 1500 bases, for example at least 1600 bases, such as at least 1700 bases, for example at least 1800 bases, such as at least 1900 bases, for example at least 2000 bases, such as at least 2100 bases, for example at least 2200 bases, such as at least 2300 bases, for example at least 2400 bases, such as at least 2500 bases, for example at least 2600 bases, such as at least 2700 bases, for example at least 2800 bases, such as at least 2900 bases, for example at least 3000 bases, such as at least 3200 bases, for example at least 3500 bases, such as at least 3800 bases, for example at least 4000 bases, such as at least 4500 bases, for example at least 5000 bases, such as at least 6000 bases.

24. The artificial chromosome according to any of the preceding claims, wherein at least one of the spacer sequences comprises between 50 and 2500 bases, such as between 100 and 2500 bases, preferably between 200 and 2300 bases, more preferably between 300 and 2100 bases, such as between 400 and 1900 bases, more preferably between 500 and 1700 bases, such as between 600 and 1500 bases, more preferably between 700 and 1400 bases.

25. The artificial chromosome according to any of the preceding claims, wherein at least one of the promoters, preferably substantially all promoters is/are an externally controllable promoter, which are functional in a host cell.

26. The artificial chromosome according to claim 25, wherein at least one of the promoters is an inducible promoter or a repressible promoter.

27. The artificial chromosome according to any of the preceding claims, comprising at least one promoter comprising both repressible and inducible elements.

28. The artificial chromosome according to any of the preceding claims, comprising at least one promoter being chemically inducible and/or repressible and/or inducible/repressible by temperature, and/or inducible/repressible according to mating type.

29. The artificial chromosome according to any of the preceding claims, comprising at least one promoter being induced by any factor selected from the group comprising carbohydrates, e.g. galactose; low inorganic phosphate levels; temperature, e.g. low or high temperature shift; metals or metal ions, e.g. copper ions; hormones, e.g. dihydrotestosterone; or deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage, e.g. developmental stage; synthetic inducers, e.g. the gal inducer.

30. The artificial chromosome according to any of the preceding claims, wherein at least one promoter is repressed by any factor selected from the group comprising carbohydrates; galactose; low inorganic phosphate levels; temperature; low or high temperature shift; metals or metal ions; copper ions; hormones; dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage; developmental stage; synthetic inducers; gal inducer; high inorganic phosphate levels; methionine; glycerol.

31. The artificial chromosome according to any of the preceding claims, wherein at least one promoter comprises a promoter selected from the group comprising ADH 1, PGK 1, GAP 491, TPI, PYK, ENO, PMA 1, PHO5, GAL 1, GAL 2, GAL 10, MET25, ADH2, MEL 1, CUP 1, HSE, AOX, MOX, SV40, CaMV, Opaque-2, GRE, ARE, PGK/ARE hybrid, CYC/GRE hybrid, TPI/α2 operator, AOX 1, MOX A.

32. The artificial chromosome according to any of the preceding claims, wherein at least one promoter is a synthetic promoter.

33. The artificial chromosome according to any of the preceding claims, wherein the terminator is capable of functioning in a host cell.

34. An artificial chromosome comprising at least a first and a second expressible nucleotide sequence under the control of a controllable promoter, the promoter of the first expressible nucleotide sequence being controllable independently from the promoter of the other expressible nucleotide sequence.

35. The artificial chromosome according to claim 1, wherein comprising at least one promoter comprising an inducible promoter or a repressible promoter.

36. The artificial chromosome according to any of the preceding claims 1 to 35, comprising at least one promoter comprising both repressible and inducible elements.

37. The artificial chromosome according to any of the preceding claims 1 to 36, comprising at least one promoter being chemically inducible and/or repressible and/or inducible/repressible by temperature, and/or inducible/repressible according to mating type.

38. The artificial chromosome according to any of the preceding claims 1 to 37, comprising at least one promoter being induced by any factor selected from the group comprising carbohydrates, e.g. galactose; low inorganic phosphase levels; temperature, e.g. low or high temperature shift; metals or metal ions, e.g. copper ions; hormones, e.g. dihydrotestosterone; or deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage, e.g. developmental stage; synthetic inducers, e.g. the gal inducer.

39. The artificial chromosome according to any of the preceding claims 1 to 38, wherein at least one promoter is repressed by any factor selected from the group comprising carbohydrates, e.g. galactose; low inorganic phosphate levels, e.g. high inorganic phosphate levels;; temperature, e.g. low or high temperature shift; metals or metal ions, e.g. copper ions; hormones, e.g. dihydrotestosterone; deoxycorticosterone; heat shock (e.g. 39°C); methanol; redox-status; growth stage, e.g. developmental stage; synthetic inducers, e.g. the gal inducer; methionine; glycerol.

40. The artificial chromosome according to any of the preceding claims 1 to 39, wherein at least one promoter comprises a promoter selected from the group comprising ADH 1, PGK 1, GAP 491, TPI, PYK, ENO, PMA 1, PHO5, GAL 1, GAL 2, GAL 10, MET25, ADH2, MEL 1, CUP 1, HSE, AOX, MOX, SV40, CaMV, Opaque-2, GRE, ARE, PGK/ARE hybrid, CYC/GRE hybrid, TPI/α2 operator, AOX 1, MOX A.

41. The artificial chromosome according to any of the preceding claims 1 to 40, wherein at least one promoter is a synthetic promoter.

42. The artificial chromosome according to any of the preceding claims 1 to 41, comprising at least 10 expressible nucleotide sequences, such as at least 15, for example at least 20, such as at least 25, for example at least 30, such as from 30 to 60 or more than 60, such as at least 75, for example at least 100, such as at least 200, for example at least 500, such as at least 750, for example at least 1000, such as at least 1500, for example at least 2000.

43. The artificial chromosome according to any of the preceding claims 1 to 42, comprising nucleotide sequences under the control of at least 3 different promoters being regulated through external manipulations, such as at least 4 different promoters, for example at least 5 different promoters, such as at least 6 different promoters, for example at least 7 different promoters, such as at least 8 different promoters, for example at least 9 different promoters, such as at least 10 different promoters, for example at least 12 different promoters, such as at least 15 different promoters, for example at least 20 different promoters, such as at least 25 different promoters, for example at least 30 different promoters, such as at least 50 different promoters or 100 different promoters.

44. The artificial chromosome according to any of the preceding claims 1 to 43, comprising at least two nucleotide sequences coding for the same peptide or two substantially identical nucleotide sequences under the control of at least 2 different promoters, such as 3 or 4 different promoters, for example at least 5 different promoters, such as at least 6 different promoters, for example at least 7 different promoters, such as at least 8 different promoters, for example at least 9 different promoters, such as at least 10 different promoters, for example at least

12 different promoters, such as at least 15 different promoters, for example at least 20 different promoters, such as at least 25 different promoters, for example at least 30 different promoters, such as at least 50 different promoters or 100 different promoters.

45. The artificial chromosome according to claim 44, comprising at least a selection of combinations of promoters and nucleotide sequences.

46. The artificial chromosome according to claim 45, whereby the selection comprises combinations from a two dimensional array of promoters and nucleotide sequences.

47. The artificial chromosome according to claim 45, whereby the selection comprises a partial or complete combination from a n-dimensional array of promoters, nucleotide sequences, spacers, terminators, and introns, wherein n is an integer from 1 to 5.

48. The artificial chromosome according to any of the preceding claims 1 to 47, wherein the artificial chromosome is selected from the group comprising a Yeast Artificial Chromosome, a mega Yeast Artificial Chromosome, a Bacterial Artificial Chromosome, a mouse artificial chromosome, a Mammalian Artificial Chromosome, an Insect Artificial Chromosome, an Avian Artificial Chromosome, a Bacteriophage Artificial Chromosome, a Baculovirus Artificial Chromosome, or a Human Artificial Chromosome.

49. The artificial chromosome according to any of the preceding claims 1 to 48, wherein the chromosome further comprises at least one selectable genetic marker, such as a recessive or a dominant marker.

50. The artificial chromosome according to claim 49, comprising at least two selectable genetic markers.

51. The artificial chromosome according to any of the preceding claims 49 to 50, wherein the at least one marker comprises a marker selected from the group comprising LEU 2, TRP 1, HIS 3, LYS 2, URA 3, ADE 2, Amyloglucosidase, β-

lactamase, CUP 1, G418^R, TUN^R, KILk1, C230, SMR1, SFA, Hygromycin^R, methotrexate^R, chloramphenicol^R, Diuron^R, Zeocin^R, Canavanine^R.

52. The artificial chromosome according to any of the preceding claims 1 to 51, being designed to minimise the level of repeat sequences occurring in the concatemer.

53. A host cell comprising at least one artificial chromosome comprising at least a first and a second expressible nucleotide sequence under the control of a controllable promoter, the promoter of the first expressible nucleotide sequence being controllable independently from the promoter of the other expressible nucleotide sequence.

54. The host cell according to claim 53, wherein the two different nucleotide sequences are from the same expression state or from at least two different expression states.

55. The cell according to claim 53, wherein the at least two different expression states represent at least two different tissues, such as at least two organs, such as at least two species, such as at least two genera.

56. The cell according to claim 55, wherein the two different species are from at least two different phylae, such as from at least two different classes, such as from at least two different divisions, more preferably from at least two different sub-kingdoms, such as from at least two different kingdoms.

57. The cell according to claim 55 or 56, wherein one species is a eukaryot and another species is a prokaryot.

58. The cell according to any of the preceding claims 53 to 57, comprising at least two sub-sets of expressible nucleotide sequences, the expressible nucleotide sequences of the first set being under the control of the same controllable promoter and the expressible nucleotide sequences of the second sub-set being under the control of another controllable promoter.

59. The cell according to claim 58, comprising at least three sub-sets of expressible nucleotide sequences, such as at least four sub-sets, for example at least five sub-sets, such as at least six sub-sets, for example at least seven sub-sets, such as at least eight sub-sets, for example at least nine sub-sets, such as at least ten sub-sets, for example a 11, 12, 15, 20, 25, 30, 50, 75 or at least 100 sub-set of expressible nucleotide sequences, each sub-set comprising a unique controllable promoter.

60. The cell according to claim 58 to 59, wherein each sub-set of nucleotide sequences comprises a random and individual selection of expressible nucleotide sequences from the same population of expressible nucleotide sequences.

61. The cell according to any of the preceding claims 53 to 60, further comprising at least one heterologous controllably expressible nucleotide sequences inserted into a native chromosome and/or being located on a plasmid and/or a cosmid, and/or a phage and/or a virus.

62. The cell according to any of claims 53 to 61, comprising a prokaryotic cell selected from the group comprising bacteria such as Escherichia coli, Bacillus subtilis, Streptomyces lividans, Streptomyces coelicolor Pseudomonas aeruginosa, Myxococcus xanthus.

63. The cell according to any of claims 53 to 621, comprising a eukaryotic cell selected from the group comprising: yeasts; filamentous ascomycetes such as Neurospora crassa and Aspergillus nidulans; plant cells such as those derived from Nicotiana and Arabidopsis; mammalian host cells such as those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa.

64. The cell according to claim 63, being a yeast cell selected from the group comprising Kluyveromyces marxianus, K. lactis, Candida utilis, Phaffia rhodozyma, Saccharomyces boulardii, Pichia pastoris, Hansenula polymorpha, Yarrowia lipolytica, Candida paraffinica, Schwanniomyces castellii, Pichia stipitis, Candida shehatae, Rhodotorula glutinis, Lipomyces lipofer,

Cryptococcos curvatus, Candida spp. (e.g. C. palmioleophila), Yarrowia lipolytica, Candida guilliermondii, Candida, Rhodotorula spp., Saccharomycopsis spp., Aureobasidium pullulans, Candida brumptii, Candida hydrocarbofumarica, Torulopsis, Candida tropicalis, Saccharomyces cerevisiae, Rhodotorula rubra, Candida flaveri, Eremothecium ashbyii, Pichia spp., Kluyveromyces, Hansenula, Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola.

65. The cell according to any of the preceding claims 53 to 64, having a mutation in a central biosynthetic pathway.

66. The cell according to claim 65, comprising a selectable genetic marker inserted on at least one artificial chromosome complementing the mutation.

67. The cell according to any of the preceding claims 53 to 66, comprising at least one selectable genetic marker inserted on at least one artificial chromosome.

68. The cell according to claim 67, comprising at least two selectable genetic markers inserted on at least one artificial chromosome.

69. The cell according to any of the preceding claims 53 to 68, wherein each artificial chromosome comprises at least one unique selectable genetic marker.

70. The cell according to claim 69, wherein each artificial chromosome comprises two unique selectable markers.

71. The cell according to claim 69, wherein all artificial chromosome comprise one common selectable marker.

72. The cell according to any of claims 53 to 69, wherein the nucleotide sequence of at least one artificial chromosome, preferably the nucleotide sequence from substantially all artificial chromosomes have been designed to minimise the level of repeat sequences in any one artificial chromosome.

73. The cell according to any of the preceding claims 53 to 72, wherein recombination within the expressible nucleotide sequence has been minimised.

74. The cell according to any of the preceding claims 53 to 73, wherein at least one artificial chromosome, preferably substantially all artificial chromosomes is/are artificial chromsome/s according to claims 1 to 52.

75. A host cell comprising at least four artificial chromosomes, wherein the four chromosomes are different.

76. The cell according to claim 75, wherein at least one artificial chromosome comprises an expressible nucleotide sequence under the control of a controllable promoter.

77. The cell according to any of the preceding claims 75 to 76, further comprising at least one heterologous controllably expressible nucleotide sequence inserted into a native chromosome and/or being located on a plasmid and/or a cosmid, and/or a phage and/or a virus.

78. The cell according to any of claims 75 to 77, comprising a prokaryotic cell selected from the group comprising bacteria such as Escherichia coli, Bacillus subtilis, Streptomyces lividans, Streptomyces coelicolor Pseudomonas aeruginosa, Myxococcus xanthus.

79. The cell according to any of claims 75 to 77, comprising a eukaryotic cell selected from the group comprising: yeasts; filamentous ascomycetes such as Neurospora crassa and Aspergillus nidulans; plant cells such as those derived from Nicotiana and Arabidopsis; mammalian host cells such as those derived from humans, monkeys and rodents, such as chinese hamster ovary (CHO) cells, NIH/3T3, COS, 293, VERO, HeLa.

80. The cell according to claim 79, being a yeast cell selected from the group comprising, Kluyveromyces marxianus, K. lactis, Candida utilis, Phaffia rhodozyma, Saccharomyces boulardii, Pichia pastoris, Hansenula polymorpha, Yarrowia lipolytica, Candida paraffinica, Schwanniomyces castellii, Pichia stipitis, Candida shehatae, Rhodotorula glutinis, Lipomyces lipofer, Cryptococcos curvatus, Candida spp. (e.g. C. palmioleophila), Yarrowia

lipolytica, Candida guilliermondii, Candida, Rhodotorula spp., Saccharomycopsis spp., Aureobasidium pullulans, Candida brumptii, Candida hydrocarbofumarica, Torulopsis, Candida tropicalis, Saccharomyces cerevisiae, Rhodotorula rubra, Candida flaveri, Eremothecium ashbyii, Pichia spp., Kluyveromyces, Hansenula, Kloeckera, Pichia, Pachysolen spp., or Torulopsis bombicola.

81. The cell according to any of the preceding claims 75 to 80, having a mutation in a central biosynthetic pathway.

82. The cell according to claim 81, comprising a selectable genetic marker inserted on an artificial chromosome complementing the mutation.

83. The cell according to any of the preceding claims 75 to 82, comprising at least one selectable genetic marker inserted on at least one artificial chromosome.

84. The cell according to claim 83, comprising at least two selectable markers inserted on at least one artificial chromosome.

85. The cell according to any of the preceding claims 75 to 84, wherein each artificial chromosome comprises at least one unique selectable genetic marker.

86. The cell according to claim 85, wherein each artificial chromosome comprises at least two unique selectable genetic markers.

87. The cell according to claim 85, wherein artificial chromosomes comprise at least one common selectable genetic marker.

88. The cell according to any of claims 75 to 85, wherein the nucleotide sequence of at least one artificial chromosome, preferably the nucleotide sequence from substantially all artificial chromosomes have been designed to minimise the level of repeat sequences in any one artificial chromosome.

89. The cell according to any of the preceding claims 75 to 88, wherein recombination within the expressible nucleotide sequence has been minimised.

90. The cell according to any of the preceding claims 75 to 89, wherein at least one artificial chromosome, preferably substantially all artificial chromosomes is/are artificial chromsome/s according to claims 1 to 52.

Fig. 1

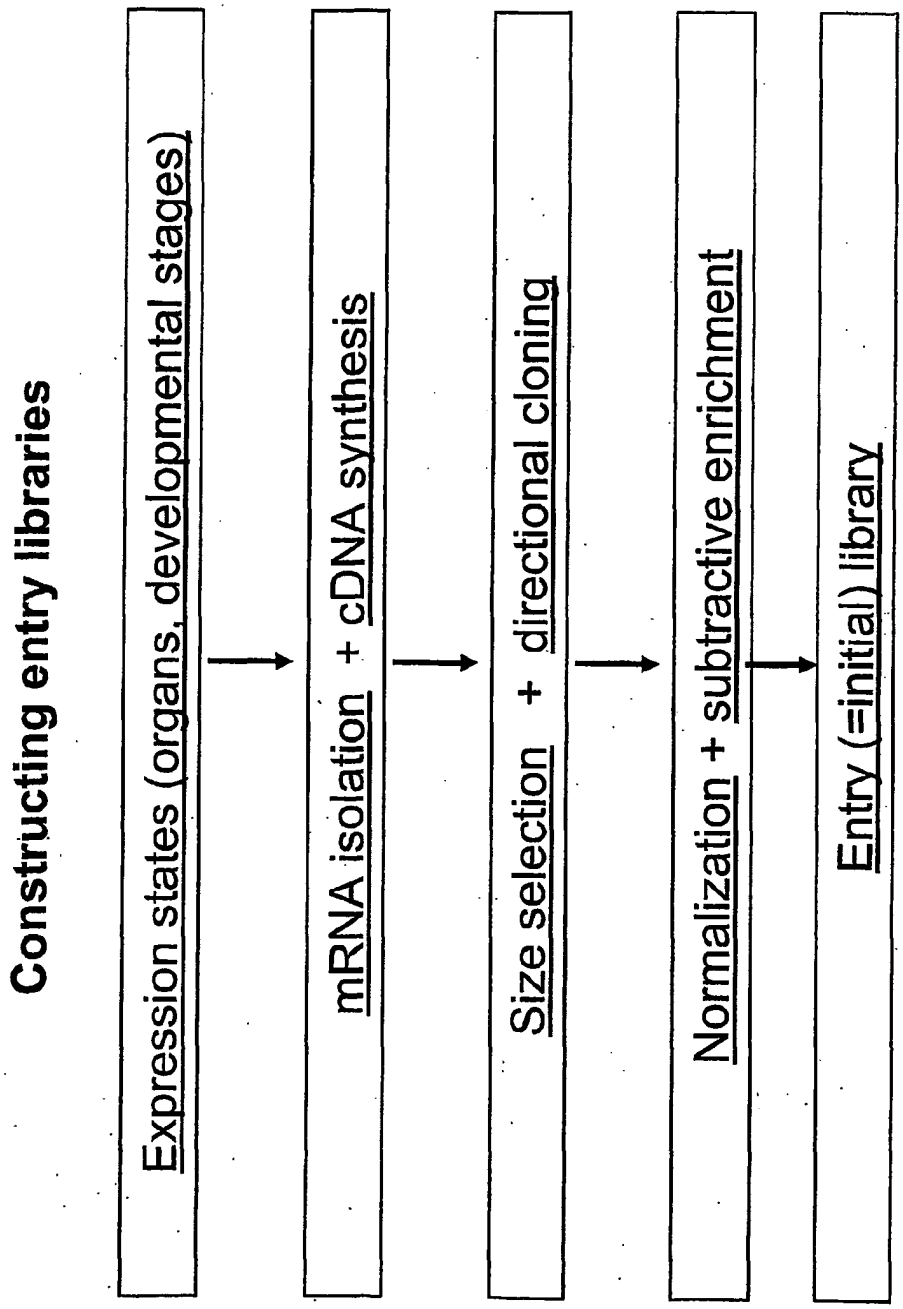

Constructing entry libraries
Expression states (organs, developmental stages)
mRNA isolation + cDNA synthesis
Size selection + directional cloning
Normalization + subtractive enrichment
Entry (=initial) library

Fig. 2a

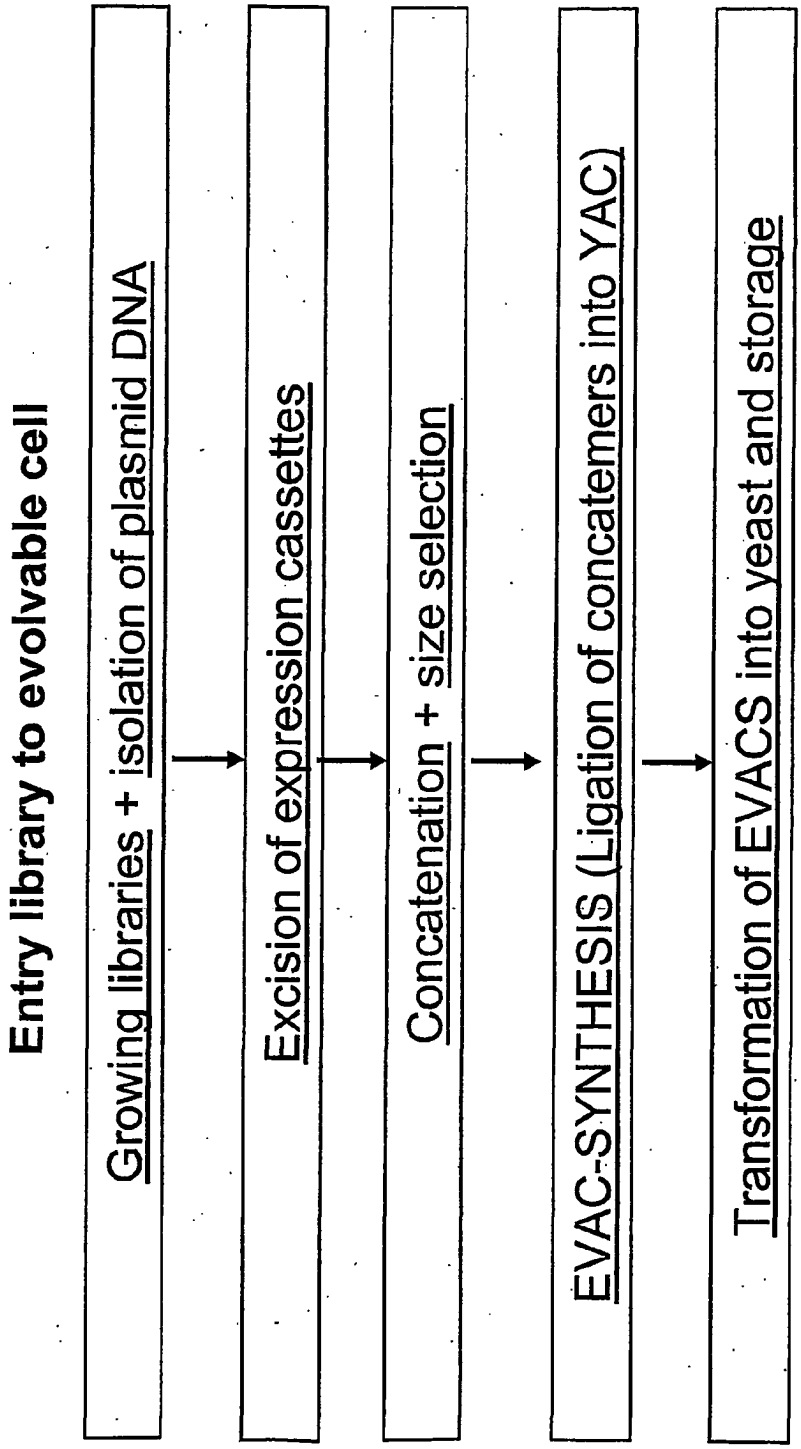

Entry library to evolvable cell
Growing libraries + isolation of plasmid DNA
Excision of expression cassettes
Concatenation + size selection
EVAC-SYNTHESIS (Ligation of concatemers into YAC)
Transformation of EVACS into yeast and storage

Fig. 2b

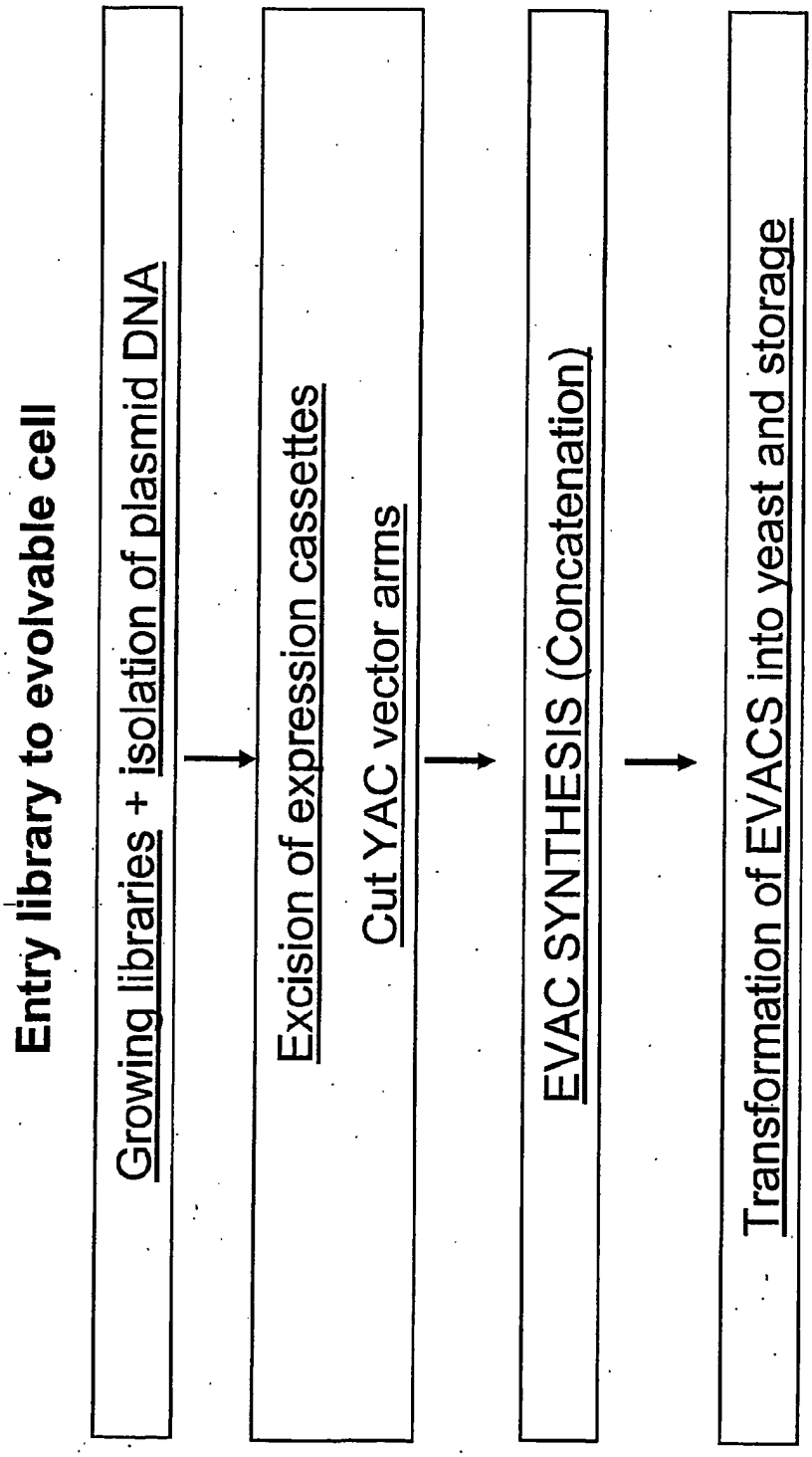

Entry library to evolvable cell
Growing libraries + isolation of plasmid DNA
Excision of expression cassettes
Cut YAC vector arms
EVAC SYNTHESIS (Concatenation)
Transformation of EVACS into yeast and storage

Fig. 3

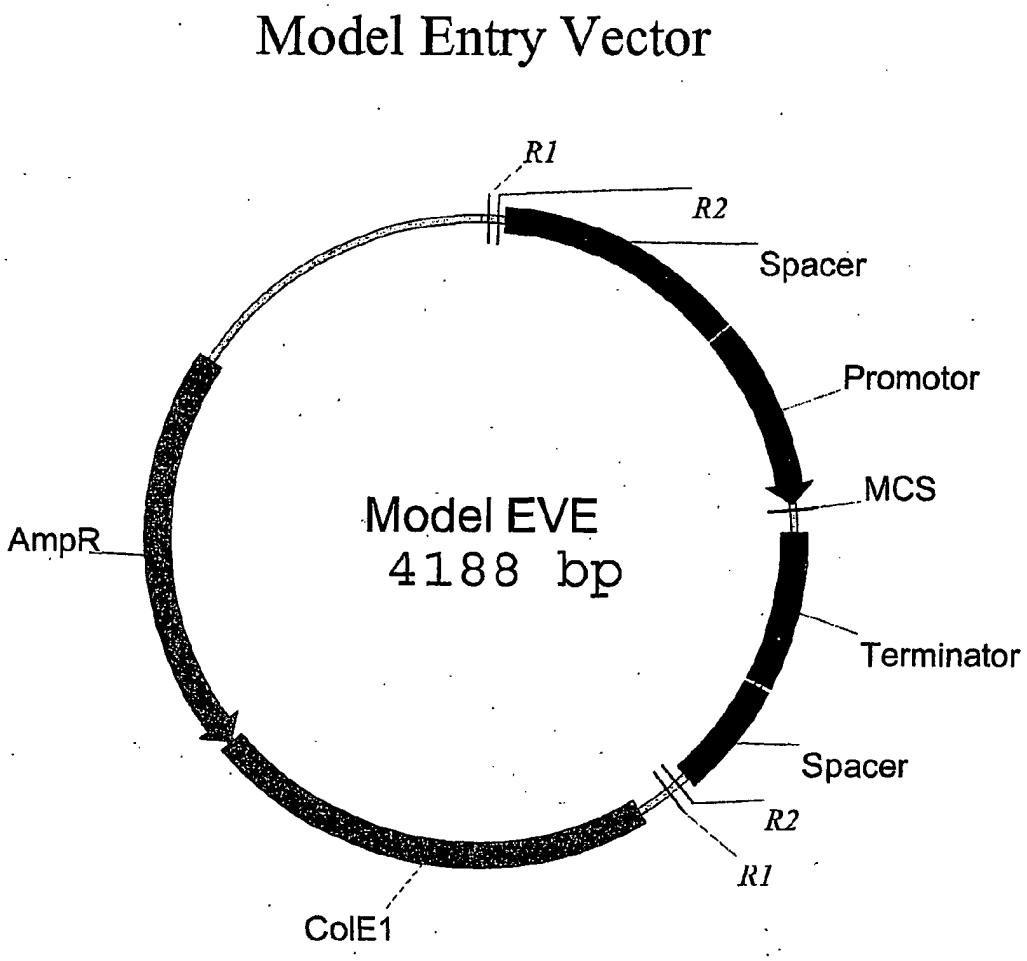

Model Entry Vector
R1
R2
Spacer
Promotor
MCS
Terminator
Spacer
R2
R1
ColE1
AmpR
Model EVE
4188 bp

Fig. 4

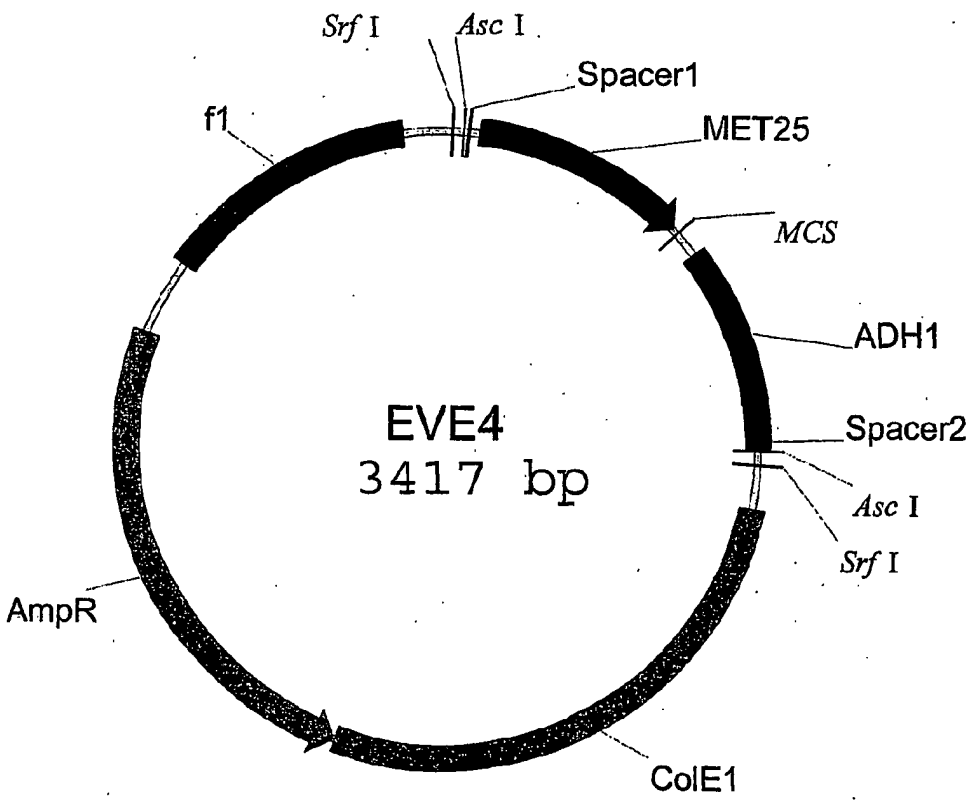

EVE4 entry vector
Srf I
Asc I
Spacer1
f1
MET25
MCS
ADH1
EVE4
3417 bp
Spacer2
Asc I
Srf I
AmpR
ColE1

Fig. 5

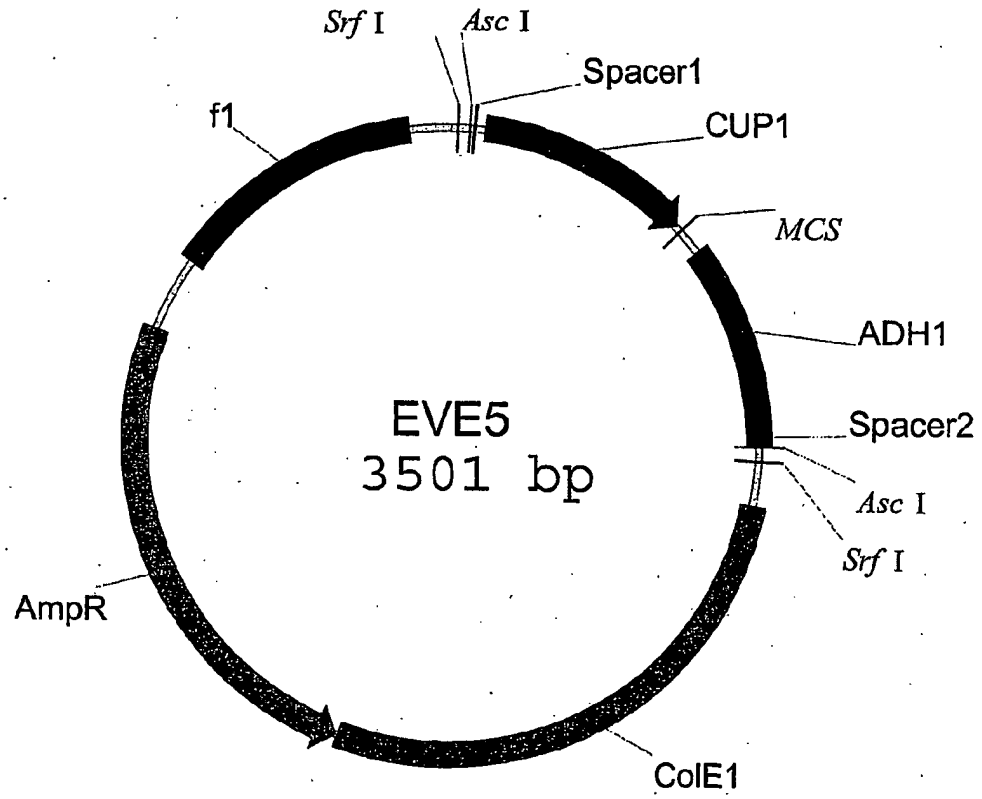

EVE5 entry vector
Srf I
Asc I
Spacer1
CUP1
f1
MCS
ADH1
EVE5
3501 bp
Spacer2
Asc I
Srf I
AmpR
ColE1

Fig. 6

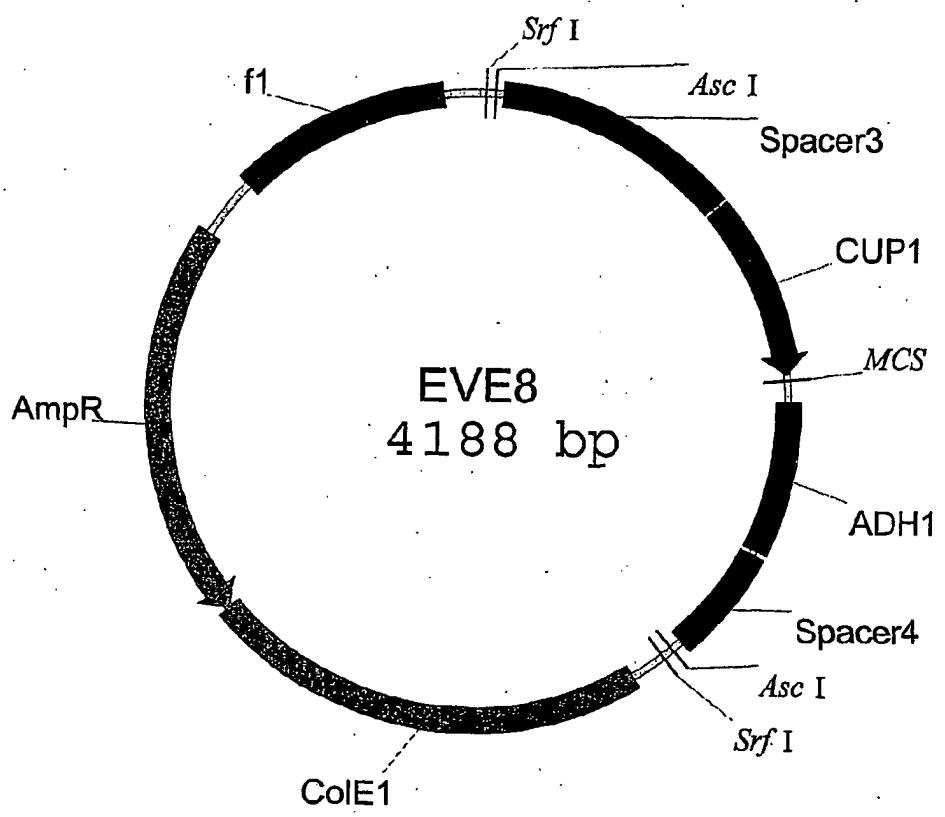

EVE8 entry vector
Srf I
Asc I
f1
Spacer3
CUP1
MCS
EVE8
4188 bp
AmpR
ADH1
Spacer4
Asc I
Srf I
ColE1

Fig. 7

pYAC4-AscI

Vector for providing EVACS arms

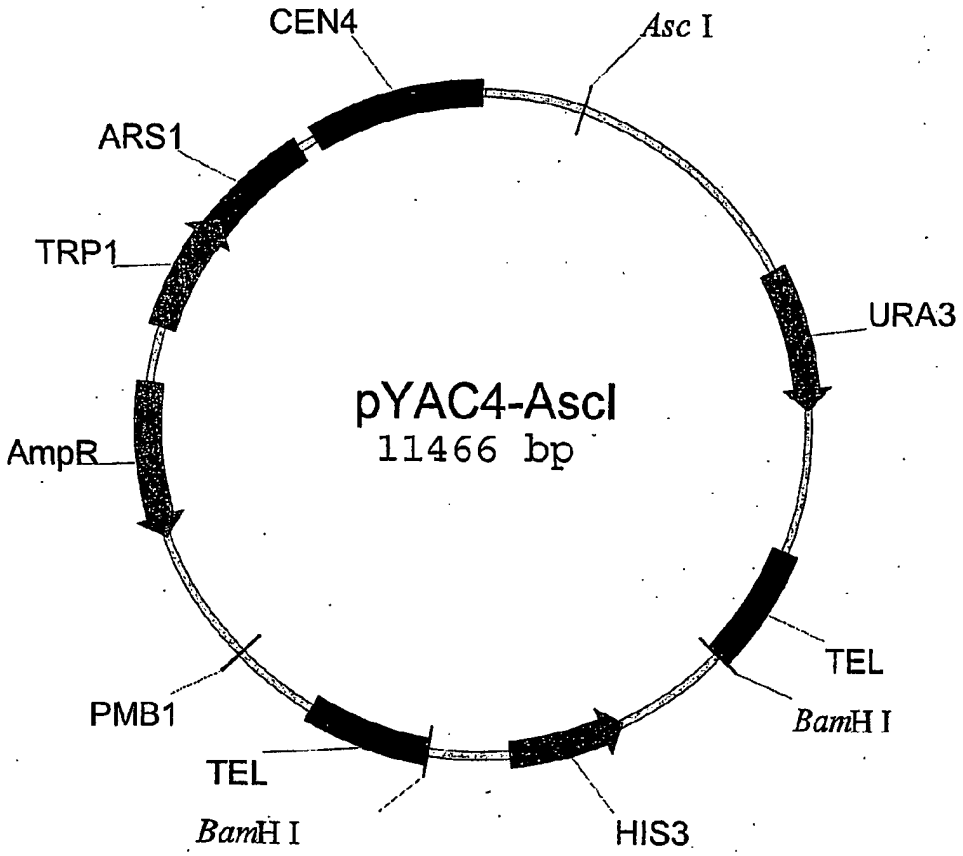


Fig. 8

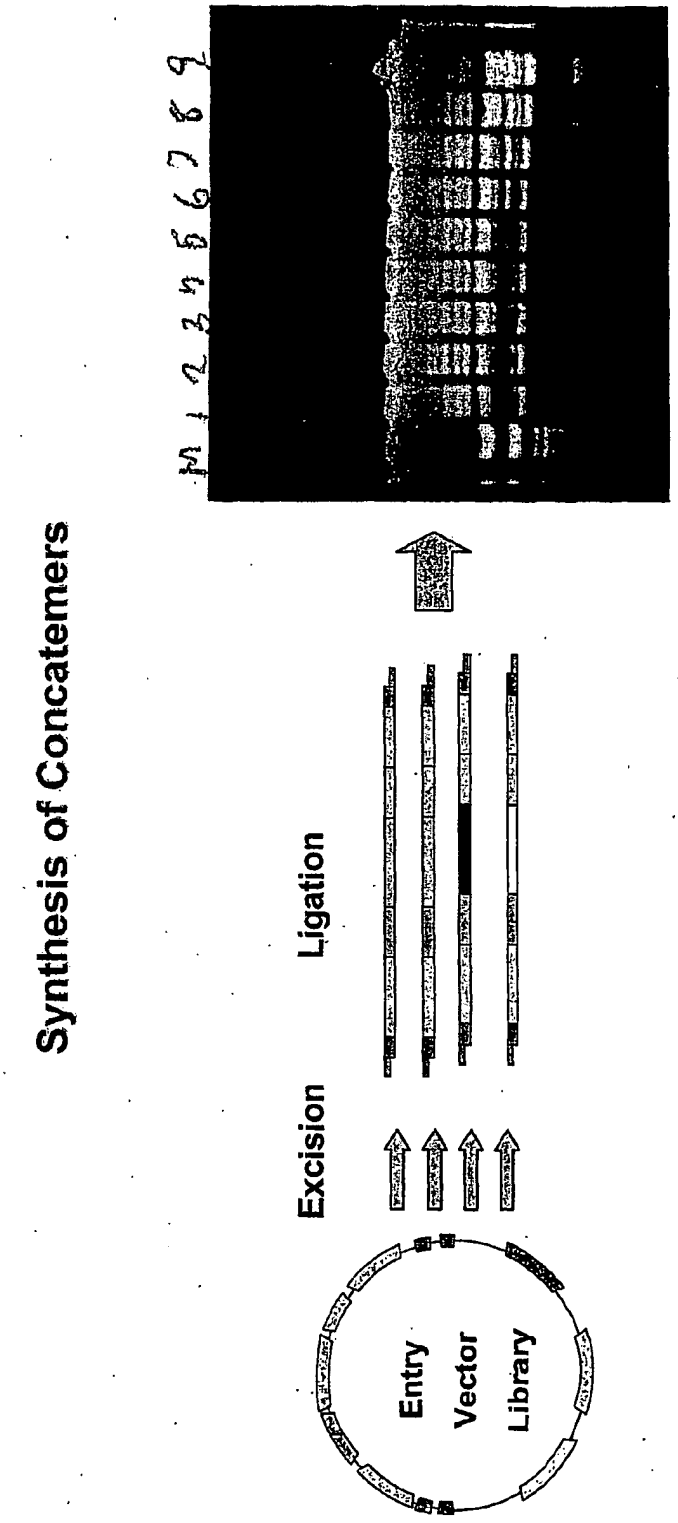
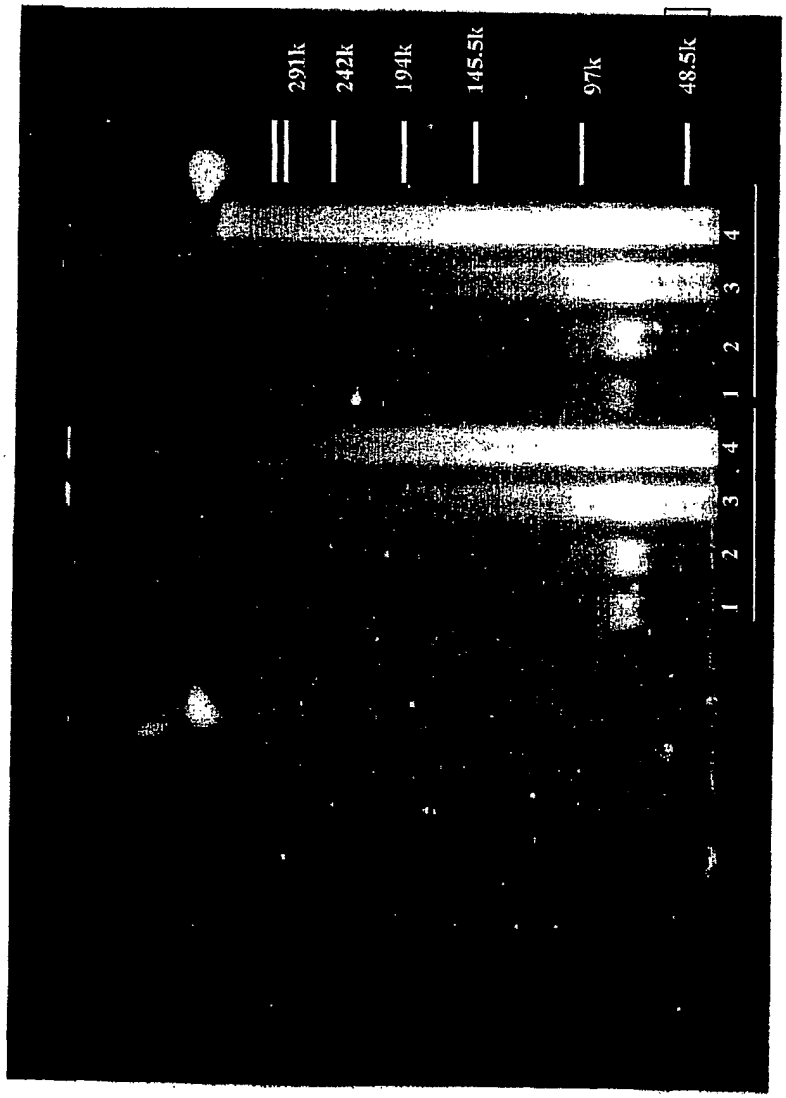

M 1 2 3 4 5 6 6 7 8 9
Synthesis of Concatemers
Ligation
Excision
Entry
Vector
Library

Fig. 9a



A: F/Y = 100, B: F/Y= 1000

1: fragment conc.= 1

2: fragment conc.= 2

3: fragment conc.= 5

4: fragment conc.= 10

3+4: same amount loaded on gel

But concentration in 4 = 2x concentration in 3

Fig. 9b

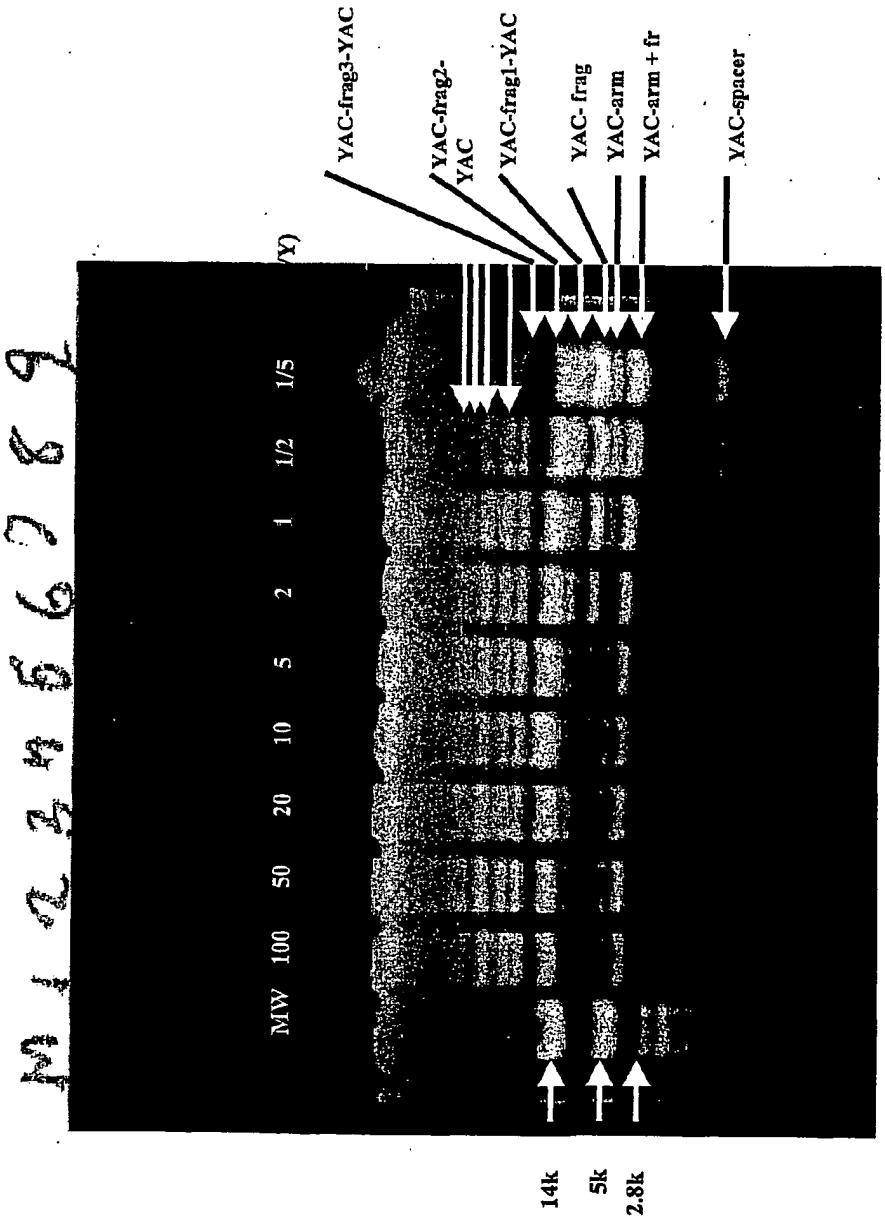

YAC-frag3-YAC
YAC-frag2-YAC
YAC-frag1-YAC
YAC- frag
YAC-arm
YAC-arm + fr
YAC-spacer
MW 100 50 20 10 5 2 1 1/2 1/5
14k
5k
2.8k

Fig. 10

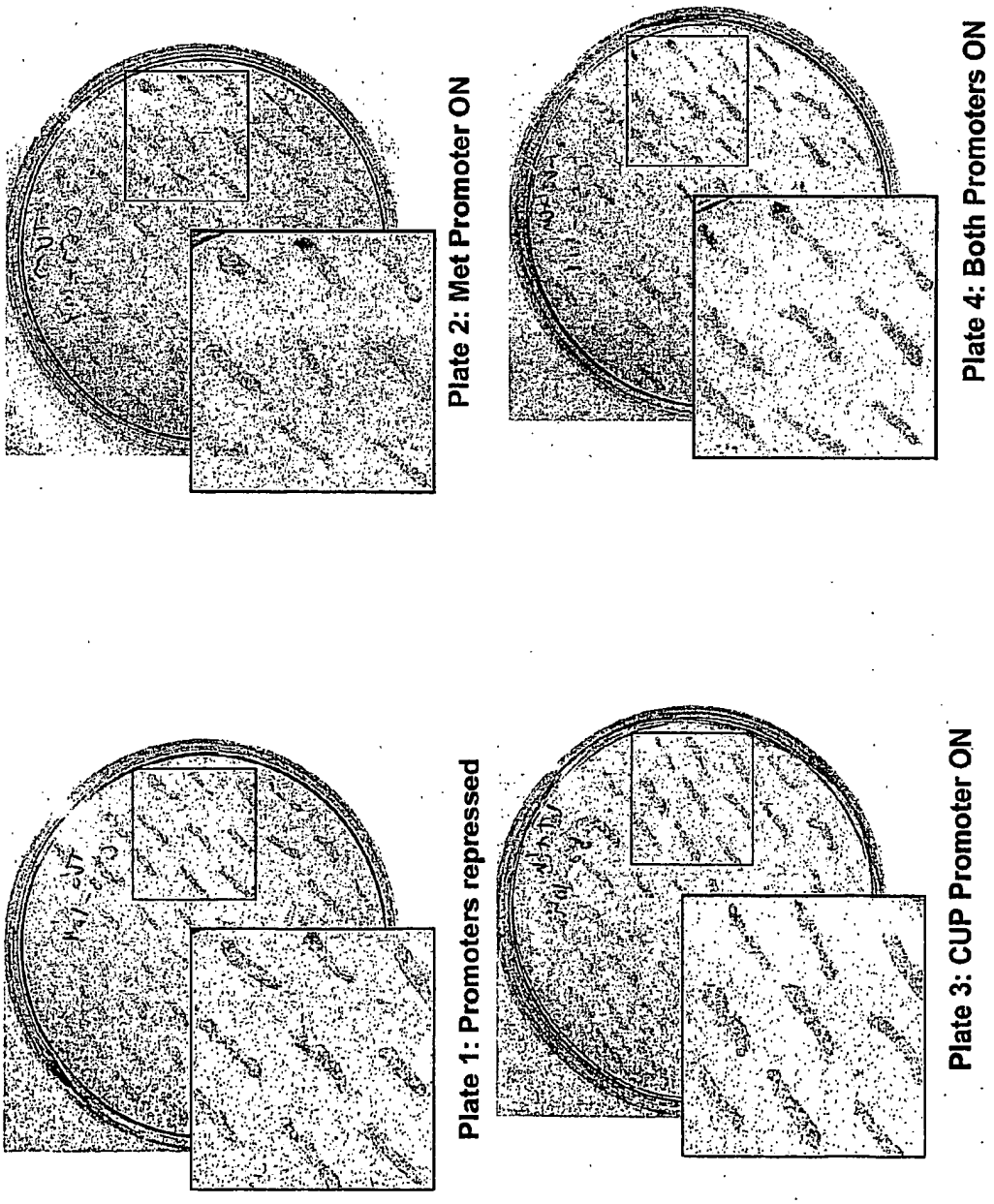
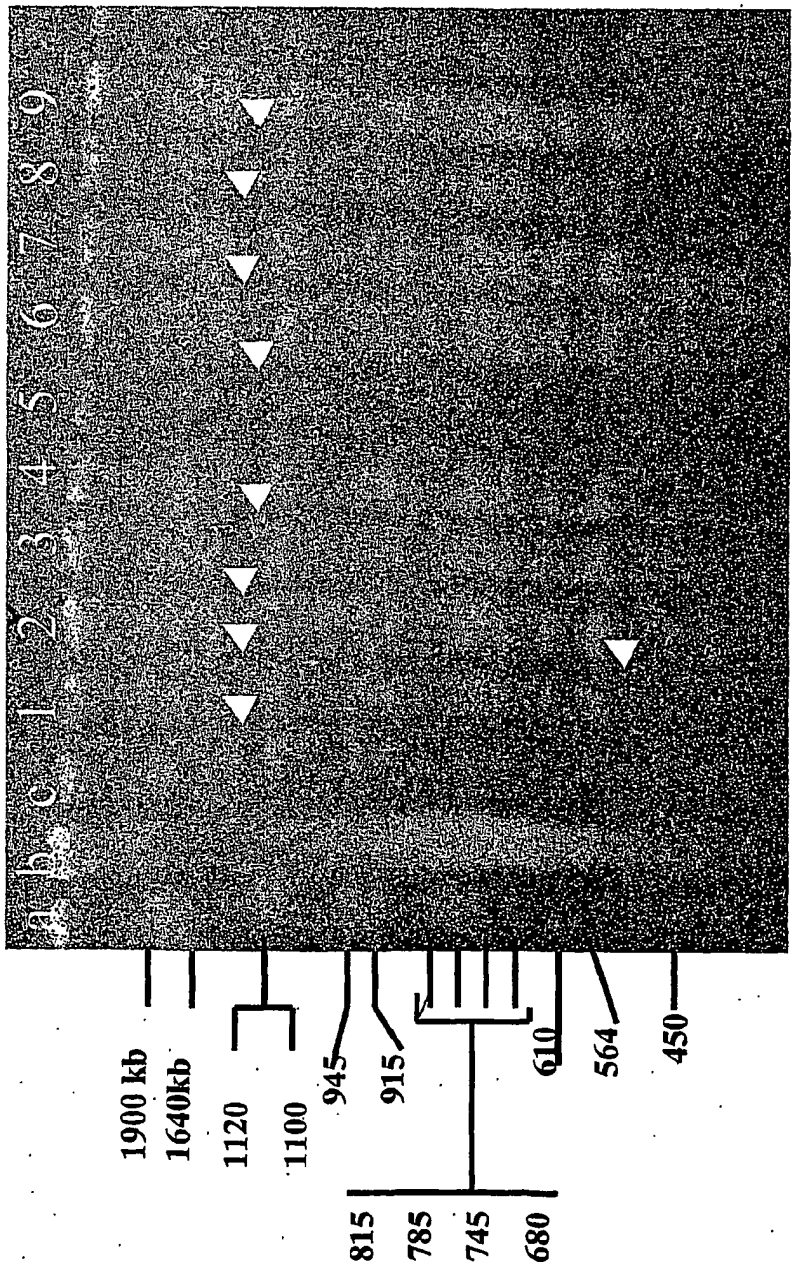


Plate 2: Met Promoter ON
Plate 4: Both Promoters ON
Plate 1: Promoters repressed
Plate 3: CUP Promoter ON

Fig. 11


a b c 1 2 3 4 5 6 7 8 9
1900 kb
1640kb
1120
1100
945
915
610
564
450
815
785
745
680

SEQUENCE LISTING

<110> Evolva Biotech AS
Goldsmith, Neil
Sørensen, Alexandra M. P.
Nielsen, Søren V.S.

<120> Artificial chromosomes comprising concatemers of expressible nucleotide sequences

<130> P 503 PC00

<150> DK PA 2001 00130
<151> 2001-01-25

<150> US 60/300,865
<151> 2001-06-27

<160> 4

<170> PatentIn version 3.1

<210> 1
<211> 3417
<212> DNA
<213> Synthetic

<220>
<221> misc_feature
<222> (1902)..(2759)
<223> Ampicillin resistance gene

<220>
<221> rep_origin
<222> (959)..(1899)
<223> ColE1

<220>
<221> misc_feature
<222> (2891)..(3347)
<223> f1-phage origin of replication

<220>
<221> terminator
<222> (495)..(823)
<223> ADH1

<220>
<221> promoter
<222> (49)..(437)
<223> Met25 promoter

<400> 1
ctgatttgcc cgggcagttc aggctcatca ggcgcgccat gcagggattc ttcggatgca 60

agggttcgaa tcccttagct ctcattattt tttgcttttt ctcttgaggt cacatgatcg 120

caaaatggca aatggcacgt gaagctgtcg atattgggga actgtggtgg ttggcaaatg 180

actaattaag ttagtcaagg cgccatcctc atgaaaactg tgtaacataa taaccgaagt 240

gtcgaaaagg tggcaccttg tccaattgaa cacgctcgat gaaaaaaata agatatatat 300

aaggttaagt aaagcgtctg ttagaaagga agtttttcct ttttcttgct ctcttgtctt 360

ttcatctact atttccttcg tgtaatacag ggtcgtcaga tacatagata caattctatt 420

acccccatcc atacaagctt ggcgccgaat tcgtcgaccc ggggatccgc ggccgcaggc 480

ctaaattgat ctagagcttt ggacttcttc gccagaggtt tggtcaagtc tccaatcaag 540

gttgtcggct tgtctacctt gccagaaatt tacgaaaaga tggaaaaggg tcaaatcgtt 600

ggtagatacg ttgttgacac ttctaaataa gcgaatttct tatgatttat gatttttatt 660

attaaataag ttataaaaaa aataagtgta tacaaatttt aaagtgactc ttaggtttta 720

aaacgaaaat tcttgttctt gagtaactct ttcctgtagg tcaggttgct ttctcaggta 780

tagcatgagg tcgctcttat tgaccacacc tctaccggca tgcccatggg ttaactgatc 840

aatgcatcct gcatggcgcg cctgatgagc ctgaactgcc cgggcaaatc agctggacgt 900

ctgcctgcat taatgaatcg gccaacgcgc ggggagaggc ggtttgcgta ttgggcgctc 960

ttccgcttcc tcgctcactg actcgctgcg ctcggtcgtt cggctgcggc gagcggtatc 1020

agctcactca aaggcggtaa tacggttatc cacagaatca ggggataacg caggaaagaa 1080

catgtgagca aaaggccagc aaaaggccag gaaccgtaaa aaggccgcgt tgctggcgtt 1140

tttccatagg ctccgccccc ctgacgagca tcacaaaaat cgacgctcaa gtcagaggtg 1200

gcgaaacccg acaggactat aaagatacca ggcgtttccc cctggaagct ccctcgtgcg 1260

ctctcctgtt ccgaccctgc cgcttaccgg atacctgtcc gcctttctcc cttcgggaag 1320

cgtggcgctt tctcatagct cacgctgtag gtatctcagt tcggtgtagg tcgttcgctc 1380

caagctgggc tgtgtgcacg aaccccccgt tcagcccgac cgctgcgcct tatccggtaa 1440

ctatcgtctt gagtccaacc cggtaagaca cgacttatcg ccactggcag cagccactgg 1500

taacaggatt agcagagcga ggtatgtagg cggtgctaca gagttcttga agtggtggcc 1560

taactacggc tacactagaa ggacagtatt tggtatctgc gctctgctga agccagttac 1620

cttcggaaaa agagttggta gctcttgatc cggcaaacaa accaccgctg gtagcggtgg 1680

ttttttttgtt tgcaagcagc agattacgcg cagaaaaaaa ggatctcaag aagatccttt 1740

gatcttttct acggggtctg acgctcagtg gaacgaaaac tcacgttaag ggattttggt 1800

catgagatta tcaaaaagga tcttcaccta gatcctttta aattaaaaat gaagttttaa 1860

atcaatctaa agtatatatg agtaaacttg gtctgacagt taccaatgct taatcagtga 1920

ggcacctatc tcagcgatct gtctatttcg ttcatccata gttgcctgac tccccgtcgt 1980

gtagataact acgatacggg agggcttacc atctggcccc agtgctgcaa tgataccgcg 2040

agacccacgc tcaccggctc cagatttatc agcaataaac cagccagccg gaagggccga 2100

gcgcagaagt ggtcctgcaa ctttatccgc ctccatccag tctattaatt gttgccggga 2160

agctagagta agtagttcgc cagttaatag tttgcgcaac gttgttgcca ttgctacagg 2220

catcgtggtg tcacgctcgt cgtttggtat ggcttcattc agctccggtt cccaacgatc 2280

aaggcgagtt acatgatccc ccatgttgtg caaaaaagcg gttagctcct tcggtcctcc 2340

gatcgttgtc agaagtaagt tggccgcagt gttatcactc atggttatgg cagcactgca 2400

taattctctt actgtcatgc catccgtaag atgcttttct gtgactggtg agtactcaac 2460

caagtcattc tgagaatagt gtatgcggcg accgagttgc tcttgcccgg cgtcaatacg 2520

ggataatacc gcgccacata gcagaacttt aaaagtgctc atcattggaa aacgttcttc 2580

ggggcgaaaa ctctcaagga tcttaccgct gttgagatcc agttcgatgt aacccactcg 2640

tgcacccaac tgatcttcag catcttttac tttcaccagc gtttctgggt gagcaaaaac 2700

aggaaggcaa aatgccgcaa aaaagggaat aagggcgaca cggaaatgtt gaatactcat 2760

actcttcctt tttcaatatt attgaagcat ttatcagggt tattgtctca tgagcggata 2820

catatttgaa tgtatttaga aaaataaaca aatagggggtt ccgcgcacat ttccccgaaa 2880

agtgccacct gacgcgccct gtagcggcgc attaagcgcg gcgggtgtgg tggttacgcg 2940

cagcgtgacc gctacacttg ccagcgccct agcgcccgct cctttcgctt tcttcccttc 3000

ctttctcgcc acgttcgccg gctttccccg tcaagctcta aatcgggggc tccctttagg 3060

gttccgattt agtgctttac ggcacctcga ccccaaaaaa cttgattagg gtgatggttc 3120

acgtagtggg ccatcgccct gatagacggt ttttcgccct ttgacgttgg agtccacgtt 3180

ctttaatagt ggactcttgt tccaaactgg aacaacactc aaccctatct cggtctattc 3240

ttttgattta taagggattt tgccgatttc ggcctattgg ttaaaaaatg agctgattta 3300

acaaaaattt aacgcgaatt ttaacaaaat attaacgctt acaatttcca ttcgccattc 3360

aggctgcgca actgttggga agggcgatcg gtgcgggcct cttcgctatt acgccag 3417

<210> 2
<211> 3501
<212> DNA
<213> Synthetic

<220>
<221> misc_feature
<222> (1986)..(2843)
<223> Ampicillin resistance gene

<220>
<221> rep_origin
<222> (1043)..(1983)
<223> ColE1

<220>
<221> misc_feature
<222> (2975)..(3431)
<223> f1-phage origin of replication

<220>
<221> terminator
<222> (579)..(907)
<223> ADH1

<220>
<221> promoter
<222> (49)..(519)
<223> Cup1 promoter

<400> 2
ctgatttgcc cgggcagttc aggctcatca ggcgcgccat gcagggataa gccgatccca 60

ttaccgacat ttgggcgcta tacgtgcata tgttcatgta tgtatctgta tttaaaacac 120

ttttgtatta ttttttcctca tatatgtgta taggtttata cggatgattt aattattact 180

tcaccaccct ttatttcagg ctgatatctt agccttgtta ctagttagaa aaagacattt 240

ttgctgtcag tcactgtcaa gagattcttt tgctggcatt tcttctagaa gcaaaaagag 300

cgatgcgtct tttccgctga accgttccag caaaaaagac taccaacgca atatggattg 360

tcagaatcat ataaaagaga agcaaataac tccttgtctt gtatcaattg cattataata 420

tcttcttgtt agtgcaatat catatagaag tcatcgaaat agatattaag aaaaacaaac 480

tgtacaatca atcaatcaat catcacataa aatgttcaaa gcttggcgcc gaattcgtcg 540

acccggggat ccgcggccgc aggcctaaat tgatctagag ctttggactt cttcgccaga 600

ggtttggtca agtctccaat caaggttgtc ggcttgtcta ccttgccaga aatttacgaa 660

aagatggaaa agggtcaaat cgttggtaga tacgttgttg acacttctaa ataagcgaat 720

ttcttatgat ttatgatttt tattattaaa taagttataa aaaaaataag tgtatacaaa 780

ttttaaagtg actcttaggt tttaaaacga aaattcttgt tcttgagtaa ctctttcctg 840

taggtcaggt tgctttctca ggtatagcat gaggtcgctc ttattgacca cacctctacc 900

ggcatgccca tgggttaact gatcaatgca tcctgcatgg cgcgcctgat gagcctgaac 960

tgcccgggca aatcagctgg acgtctgcct gcattaatga atcggccaac gcgcggggag 1020

aggcggtttg cgtattgggc gctcttccgc ttcctcgctc actgactcgc tgcgctcggt 1080

cgttcggctg cggcgagcgg tatcagctca ctcaaaggcg gtaatacggt tatccacaga 1140

atcaggggat aacgcaggaa agaacatgtg agcaaaaggc cagcaaaagg ccaggaaccg 1200

taaaaaggcc gcgttgctgg cgtttttcca taggctccgc cccctgacg agcatcacaa 1260

aaatcgacgc tcaagtcaga ggtggcgaaa cccgacagga ctataaagat accaggcgtt 1320

tccccctgga agctccctcg tgcgctctcc tgttccgacc ctgccgctta ccggatacct 1380

gtccgccttt ctcccttcgg gaagcgtggc gctttctcat agctcacgct gtaggtatct 1440

cagttcggtg taggtcgttc gctccaagct gggctgtgtg cacgaacccc ccgttcagcc 1500

cgaccgctgc gccttatccg gtaactatcg tcttgagtcc aacccggtaa gacacgactt 1560

atcgccactg gcagcagcca ctggtaacag gattagcaga gcgaggtatg taggcggtgc 1620

tacagagttc ttgaagtggt ggcctaacta cggctacact agaaggacag tatttggtat 1680

ctgcgctctg ctgaagccag ttaccttcgg aaaaagagtt ggtagctctt gatccggcaa 1740

acaaaccacc gctggtagcg gtggtttttt tgtttgcaag cagcagatta cgcgcagaaa 1800

aaaaggatct caagaagatc ctttgatctt ttctacgggg tctgacgctc agtggaacga 1860

aaactcacgt taagggattt tggtcatgag attatcaaaa aggatcttca cctagatcct 1920

tttaaattaa aaatgaagtt ttaaatcaat ctaaagtata tatgagtaaa cttggtctga 1980

cagttaccaa tgcttaatca gtgaggcacc tatctcagcg atctgtctat ttcgttcatc 2040

catagttgcc tgactccccg tcgtgtagat aactacgata cgggagggct taccatctgg 2100

ccccagtgct gcaatgatac cgcgagaccc acgctcaccg gctccagatt tatcagcaat 2160

aaaccagcca gccggaaggg ccgagcgcag aagtggtcct gcaactttat ccgcctccat 2220

ccagtctatt aattgttgcc gggaagctag agtaagtagt tcgccagtta atagtttgcg 2280

caacgttgtt gccattgcta caggcatcgt ggtgtcacgc tcgtcgtttg gtatggcttc 2340

attcagctcc ggttcccaac gatcaaggcg agttacatga tcccccatgt tgtgcaaaaa 2400

agcggttagc tccttcggtc ctccgatcgt tgtcagaagt aagttggccg cagtgttatc 2460

actcatggtt atggcagcac tgcataattc tcttactgtc atgccatccg taagatgctt 2520

ttctgtgact ggtgagtact caaccaagtc attctgagaa tagtgtatgc ggcgaccgag 2580

ttgctcttgc ccggcgtcaa tacgggataa taccgcgcca catagcagaa ctttaaaagt 2640

gctcatcatt ggaaaacgtt cttcggggcg aaaactctca aggatcttac cgctgttgag 2700

atccagttcg atgtaaccca ctcgtgcacc caactgatct tcagcatctt ttactttcac 2760

cagcgtttct gggtgagcaa aaacaggaag gcaaaatgcc gcaaaaaagg gaataagggc 2820

gacacggaaa tgttgaatac tcatactctt cctttttcaa tattattgaa gcatttatca 2880

gggttattgt ctcatgagcg gatacatatt tgaatgtatt tagaaaaata aacaaatagg 2940

ggttccgcgc acatttcccc gaaaagtgcc acctgacgcg ccctgtagcg gcgcattaag 3000

cgcggcgggt gtggtggtta cgcgcagcgt gaccgctaca cttgccagcg ccctagcgcc 3060

cgctcctttc gctttcttcc cttcctttct cgccacgttc gccggctttc cccgtcaagc 3120

tctaaatcgg gggctccctt tagggttccg atttagtgct ttacggcacc tcgaccccaa 3180

aaaacttgat tagggtgatg gttcacgtag tgggccatcg ccctgataga cggtttttcg 3240

ccctttgacg ttggagtcca cgttctttaa tagtggactc ttgttccaaa ctggaacaac 3300

actcaaccct atctcggtct attcttttga tttataaggg attttgccga tttcggccta 3360

ttggttaaaa aatgagctga tttaacaaaa atttaacgcg aattttaaca aaatattaac 3420

gcttacaatt tccattcgcc attcaggctg cgcaactgtt gggaagggcg atcggtgcgg 3480

gcctcttcgc tattacgcca g 3501

<210> 3
<211> 4188
<212> DNA
<213> Synthetic

<220>
<221> misc_feature
<222> (2673)..(3530)
<223> Ampicillin resistance gene

<220>
<221> rep_origin
<222> (1730)..(2670)
<223> ColE1

<220>
<221> misc_feature
<222> (3662)..(4118)
<223> f1-phage origin of replication

<220>
<221> terminator
<222> (1027)..(1355)
<223> ADH1

<220>
<221> promoter
<222> (582)..(969)
<223> Met25 promoter

<220>
<221> misc_feature
<222> (1365)..(1603)
<223> ARS1 (autonomous replicating sequence) for Yeast replication

<220>
<221> misc_feature
<222> (49)..(574)
<223> lambda spacer DNA (22428-22923)

<400> 3
ctgatttgcc cgggcagttc aggctcatca ggcgcgccat gcagggattc tggaaattgc 60
aacgaaggaa gaaacctcgt tgctggaagc ctggaagaag tatcgggtgt tgctgaaccg 120
tgttgataca tcaactgcac ctgatattga gtggcctgct gtccctgtta tggagtaatc 180
gttttgtgat atgccgcaga aacgttgtat gaaataacgt tctgcggtta gttagtatat 240
tgtaaagctg agtattggtt tatttggcga ttattatctt caggagaata atggaagttc 300
tatgactcaa ttgttcatag tgtttacatc accgccaatt gcttttaaga ctgaacgcat 360
gaaatatggt ttttcgtcat gttttgagtc tgctgttgat atttctaaag tcggtttttt 420
ttcttcgttt tctctaacta ttttccatga aatacatttt tgattattat ttgaatcaat 480
tccaattacc tgaagtcttt catctataat tggcattgta tgtattggtt tattggagta 540
gatgcttgct tttctgagcc atagctctga tatcagatct tcttcggatg caagggttcg 600
aatcccttag ctctcattat tttttgcttt ttctcttgag gtcacatgat cgcaaaatgg 660
caaatggcac gtgaagctgt cgatattggg gaactgtggt ggttggcaaa tgactaatta 720
agttagtcaa ggcgccatcc tcatgaaaac tgtgtaacat aataaccgaa gtgtcgaaaa 780
ggtggcacct tgtccaattg aacacgctcg atgaaaaaaa taagatatat ataaggttaa 840
gtaaagcgtc tgttagaaag gaagtttttc ctttttcttg ctctcttgtc ttttcatcta 900
ctatttcctt cgtgtaatac agggtcgtca gatacataga tacaattcta ttacccccat 960
ccatacaagc ttggcgccga attcgtcgac ccggggatcc gcggccgcag gcctaaattg 1020
atctagagct ttggacttct tcgccagagg tttggtcaag tctccaatca aggttgtcgg 1080
cttgtctacc ttgccagaaa tttacgaaaa gatggaaaag ggtcaaatcg ttggtagata 1140
cgttgttgac acttctaaat aagcgaattt cttatgattt atgattttta ttattaaata 1200
agttataaaa aaaataagtg tatacaaatt ttaaagtgac tcttaggttt taaaacgaaa 1260
attcttgttc ttgagtaact ctttcctgta ggtcaggttg ctttctcagg tatagcatga 1320
ggtcgctctt attgaccaca cctctaccgg catgcccatg ggttcttttg aaaagcaagc 1380
ataaaagatc taaacataaa atctgtaaaa taacaagatg taaagataat gctaaatcat 1440
ttggcttttt gattgattgt acaggaaaat atacatcgca gggggttgac ttttaccatt 1500
tcaccgcaat ggaatcaaac ttgttgaaga gaatgttcac aggcgcatac gctacaatga 1560
cccgattctt gctagccttt tctcggtctt gcaaacaacc gccaactgat caatgcatcc 1620
tgcatggcgc gcctgatgag cctgaactgc ccgggcaaat cagctggacg tctgcctgca 1680

ttaatgaatc ggccaacgcg cggggagagg cggtttgcgt attgggcgct cttccgcttc 1740
ctcgctcact gactcgctgc gctcggtcgt tcggctgcgg cgagcggtat cagctcactc 1800
aaaggcggta atacggttat ccacagaatc aggggataac gcaggaaaga acatgtgagc 1860
aaaaggccag caaaaggcca ggaaccgtaa aaaggccgcg ttgctggcgt ttttccatag 1920
gctccgcccc cctgacgagc atcacaaaaa tcgacgctca agtcagaggt ggcgaaaccc 1980
gacaggacta taaagatacc aggcgtttcc ccctggaagc tccctcgtgc gctctcctgt 2040
tccgaccctg ccgcttaccg gatacctgtc cgcctttctc ccttcgggaa gcgtggcgct 2100
ttctcatagc tcacgctgta ggtatctcag ttcggtgtag gtcgttcgct ccaagctggg 2160
ctgtgtgcac gaacccccgg ttcagcccga ccgctgcgcc ttatccggta actatcgtct 2220
tgagtccaac ccggtaagac acgacttatc gccactggca gcagccactg gtaacaggat 2280
tagcagagcg aggtatgtag gcggtgctac agagttcttg aagtggtggc ctaactacgg 2340
ctacactaga aggacagtat ttggtatctg cgctctgctg aagccagtta ccttcggaaa 2400
aagagttggt agctcttgat ccggcaaaca aaccaccgct ggtagcggtg gttttttttgt 2460
ttgcaagcag cagattacgc gcagaaaaaa aggatctcaa gaagatcctt tgatcttttc 2520
tacggggtct gacgctcagt ggaacgaaaa ctcacgttaa gggattttgg tcatgagatt 2580
atcaaaaagg atcttcacct agatcctttt aaattaaaaa tgaagtttta aatcaatcta 2640
aagtatatat gagtaaactt ggtctgacag ttaccaatgc ttaatcagtg aggcacctat 2700
ctcagcgatc tgtctatttc gttcatccat agttgcctga ctccccgtcg tgtagataac 2760
tacgatacgg gagggcttac catctggccc cagtgctgca atgataccgc gagacccacg 2820
ctcaccggct ccagatttat cagcaataaa ccagccagcc ggaagggccg agcgcagaag 2880
tggtcctgca actttatccg cctccatcca gtctattaat tgttgccggg aagctagagt 2940
aagtagttcg ccagttaata gtttgcgcaa cgttgttgcc attgctacag gcatcgtggt 3000
gtcacgctcg tcgtttggta tggcttcatt cagctccggt tcccaacgat caaggcgagt 3060
tacatgatcc cccatgttgt gcaaaaaagc ggttagctcc ttcggtcctc cgatcgttgt 3120
cagaagtaag ttggccgcag tgttatcact catggttatg gcagcactgc ataattctct 3180
tactgtcatg ccatccgtaa gatgcttttc tgtgactggt gagtactcaa ccaagtcatt 3240
ctgagaatag tgtatgcggc gaccgagttg ctcttgcccg gcgtcaatac gggataatac 3300
cgcgccacat agcagaactt taaaagtgct catcattgga aaacgttctt cggggcgaaa 3360
actctcaagg atcttaccgc tgttgagatc cagttcgatg taacccactc gtgcacccaa 3420
ctgatcttca gcatcttta ctttcaccag cgtttctggg tgagcaaaaa caggaaggca 3480
aaatgccgca aaaaagggaa taagggcgac acggaaatgt tgaatactca tactcttcct 3540

ttttcaatat tattgaagca tttatcaggg ttattgtctc atgagcggat acatatttga 3600

atgtatttag aaaaataaac aaataggggt tccgcgcaca tttccccgaa aagtgccacc 3660

tgacgcgccc tgtagcggcg cattaagcgc ggcgggtgtg gtggttacgc gcagcgtgac 3720

cgctacactt gccagcgccc tagcgcccgc tcctttcgct ttcttccctt cctttctcgc 3780

cacgttcgcc ggctttcccc gtcaagctct aaatcggggg ctccctttag ggttccgatt 3840

tagtgcttta cggcacctcg accccaaaaa acttgattag ggtgatggtt cacgtagtgg 3900

gccatcgccc tgatagacgg tttttcgccc tttgacgttg gagtccacgt tctttaatag 3960

tggactcttg ttccaaactg gaacaacact caaccctatc tcggtctatt cttttgattt 4020

ataagggatt ttgccgattt cggcctattg gttaaaaaat gagctgattt aacaaaaatt 4080

taacgcgaat tttaacaaaa tattaacgct tacaatttcc attcgccatt caggctgcgc 4140

aactgttggg aagggcgatc ggtgcgggcc tcttcgctat tacgccag 4188

<210> 4
<211> 11466
<212> DNA
<213> Synthetic

<220>
<221> misc_feature
<222> (3560)..(4247)
<223> Tetrahymena thermophila macronuclear telomere

<220>
<221> misc_feature
<222> (6024)..(6711)
<223> Tetrahymena thermophila macronuclear telomere

<220>
<221> misc_feature
<222> (9644)..(10388)
<223> Autonomous replicating sequence

<220>
<221> misc_feature
<222> (10488)..(11465)
<223> Centromere IV

<220>
<221> rep_origin
<222> (7198)..(7198)
<223> Origin of replication, PMB1

<220>
<221> misc_feature
<222> (1962)..(2765)
<223> URA3, orotidine-5'-phosphate decarboxylase coding sequence

<220>
<221> misc_feature
<222> (4893)..(5552)
<223> HIS3, imidazoleglycerolphosphate dehydratase, coding sequence

<220>
<221> misc_feature
<222> (7956)..(8816)
<223> AP(R), beta-lactamase, ampR ampicillin resistance, coding sequenc
e

<220>
<221> misc_feature
<222> (9129)..(9803)
<223> TRP1, phosphoribosylanthranilate isomerase, coding sequence

<400> 4
ttctcatgtt tgacagctta tcatcgataa gctttaatgc ggtagtttat cacagttaaa 60

ttgctaacgc agtcaggcac cgtgtatgaa atctaacaat gcgctcatcg tcatcctcgg 120

caccgtcacc ctggatgctg taggcatagg cttggttatg ccggtactgc cgggcctctt 180

gcgggatatc gtccattccg acagcatcgc cagtcactat ggcgtgctgc tagcgctata 240

tgcgttgatg caatttctat gcgcacccgt tctcggagca ctgtccgacc gctttggccg 300

ccgcccagtc ctgctcgctt cgctacttgg agccactatc gactacgcga tcatggcgac 360

cacacccgtc ctgtggatca attcccttta gtataaattt cactctgaac catcttggaa 420

ggaccggtaa ttatttcaaa tctcttttttc aattgtatat gtgttatgtt atgtagtata 480

ctctttcttc aacaattaaa tactctcggt agccaagttg gtttaaggcg caagacttta 540

atttatcact acggaattgg cgcgccaatt ccgtaatctt gagatcgggc gttcgatcgc 600

cccgggagat ttttttgttt tttatgtctt ccattcactt cccagacttg caagttgaaa 660

tatttctttc aagggaattg atcctctacg ccggacgcat cgtggccggc atcaccggcg 720

ccacaggtgc ggttgctggc gcctatatcg ccgacatcac cgatggggaa gatcgggctc 780

gccacttcgg gctcatgagc gcttgtttcg gcgtgggtat ggtggcaggc cccgtggccg 840

ggggactgtt gggcgccatc tccttgcatg caccattcct tgcggcggcg gtgctcaacg 900

gcctcaacct actactgggc tgcttcctaa tgcaggagtc gcataaggga gagcgtcgac 960

cgatgccctt gagagccttc aacccagtca gctccttccg gtgggcgcgg ggcatgacta 1020

tcgtcgccgc acttatgact gtcttcttta tcatgcaact cgtaggacag gtgccggcag 1080

cgctctgggt cattttcggc gaggaccgct ttcgctggag cgcgacgatg atcggcctgt 1140

cgcttgcggt attcggaatc ttgcacgccc tcgctcaagc cttcgtcact ggtcccgcca 1200

ccaaacgttt cggcgagaag caggccatta tcgccggcat ggcggccgac gcgctgggct 1260

acgtcttgct ggcgttcgcg acgcgaggct ggatggcctt ccccattatg attcttctcg 1320

cttccggcgg catcgggatg cccgcgttgc aggccatgct gtccaggcag gtagatgacg 1380

accatcaggg acagcttcaa ggatcgctcg cggctcttac cagcctaact tcgatcactg 1440

gaccgctgat cgtcacggcg atttatgccg cctcggcgag cacatggaac gggttggcat 1500

ggattgtagg cgccgcccta taccttgtct gcctcccccgc gttgcgtcgc ggtgcatgga 1560

gccgggccac ctcgacctga atggaagccg gcggcacctc gctaacggat tcaccactcc 1620

aagaattgga gccaatcaat tcttgcggag aactgtgaat gcgcaaacca acccttggca 1680

gaacatatcc atcgcgtccg ccatctccag cagccgcacg cggcgcatcc ccccccccct 1740

ttcaattcaa ttcatcattt ttttttatt ctttttttg atttcggttt ctttgaaatt 1800

tttttgattc ggtaatctcc gaacagaagg aagaacgaag gaaggagcac agacttagat 1860

tggtatatat acgcatatgt agtgttgaag aaacatgaaa ttgcccagta ttcttaaccc 1920

aactgcacag aacaaaaacc tgcaggaaac gaagataaat catgtcgaaa gctacatata 1980

aggaacgtgc tgctactcat cctagtcctg ttgctgccaa gctatttaat atcatgcacg 2040

aaaagcaaac aaacttgtgt gcttcattgg atgttcgtac caccaaggaa ttactggagt 2100

tagttgaagc attaggtccc aaaatttgtt tactaaaaac acatgtggat atcttgactg 2160

atttttccat ggagggcaca gttaagccgc taaaggcatt atccgccaag tacaattttt 2220

tactcttcga agacagaaaa tttgctgaca ttggtaatac agtcaaattg cagtactctg 2280

cgggtgtata cagaatagca gaatgggcag acattacgaa tgcacacggt gtggtgggcc 2340

caggtattgt tagcggtttg aagcaggcgg cagaagaagt aacaaaggaa cctagaggcc 2400

ttttgatgtt agcagaattg tcatgcaagg gctccctatc tactggagaa tatactaagg 2460

gtactgttga cattgcgaag agcgacaaag attttgttat cggctttatt gctcaaagag 2520

acatgggtgg aagagatgaa ggttacgatt ggttgattat gacacccggt gtgggtttag 2580

atgacaaggg agacgcattg ggtcaacagt atagaaccgt ggatgatgtg gtctctacag 2640

gatctgacat tattattgtt ggaagaggac tatttgcaaa gggaagggat gctaaggtag 2700

agggtgaacg ttacagaaaa gcaggctggg aagcatattt gagaagatgc ggccagcaaa 2760

actaaaaaac tgtattataa gtaaatgcat gtatactaaa ctcacaaatt agagcttcaa 2820

tttaattata tcagttatta ctcgggcgta atgattttta taatgacgaa aaaaaaaaaa 2880

ttggaaagaa aaggggggg gggcagcgtt gggtcctggc cacgggtgcg catgatcgtg 2940

ctcctgtcgt tgaggacccg gctaggctgg cggggttgcc ttactggtta gcagaatgaa 3000

tcaccgatac gcgagcgaac gtgaagcgac tgctgctgca aaacgtctgc gacctgagca 3060

acaacatgaa tggtcttcgg tttccgtgtt tcgtaaagtc tggaaacgcg gaagtcagcg 3120
ccctgcacca ttatgttccg gatctgcatc gcaggatgct gctggctacc ctgtggaaca 3180
cctacatctg tattaacgaa gcgctggcat tgaccctgag tgatttttct ctggtcccgc 3240
cgcatccata ccgccagttg tttaccctca caacgttcca gtaaccgggc atgttcatca 3300
tcagtaaccc gtatcgtgag catcctctct cgtttcatcg gtatcattac ccccatgaac 3360
agaaattccc ccttacacgg aggcatcaag tgaccaaaca ggaaaaaacc gcccttaaca 3420
tggcccgctt tatcagaagc cagacattaa cgcttctgga gaaactcaac gagctggacg 3480
cggatgaaca ggcagacatc tgtgaatcgc ttcacgacca cgctgatgag ctttaccgca 3540
gccctcgagg gataagcttc atttttagat aaaatttatt aatcatcatt aatttcttga 3600
aaaacatttt atttattgat cttttataac aaaaaaccct tctaaaagtt tatttttgaa 3660
tgaaaaactt ataaaaattt atgaaaacta caaaaaataa aatttttaat taaaataatt 3720
ttgataagaa cttcaatctt tgactagcta gcttagtcat ttttgagatt taattaatat 3780
tttatgttta ttcatatata aactattcaa aatattatag aatttaaaca ttttaacatc 3840
ttaatcattc ataaataact aaaaatcaaa gtattacatc aataaataac ttttactcaa 3900
tgtcaaagaa ttattggggt tggggttggg gttggggttg gggttggggt tggggttggg 3960
gttggggttg gggttggggt tggggttggg gttggggttg gggttggggt tggggttggg 4020
gttggggttg gggttggggt tggggttggg gttggggttg gggttggggt tggggttggg 4080
gttggggttg gggttggggt tggggttggg gttggggttg gggttggggt tggggttggg 4140
gttggggttg gggttggggt tggggttggg gttggggttg gggtgggaaa acagcattca 4200
ggtattagaa gaatatcctg attcaggtga aaatattgtt gatgcgcggg atcctcgggg 4260
acaccaaata tggcgatctc ggccttttcg tttcttggag ctgggacatg tttgccatcg 4320
atccatctac caccagaacg gccgttagat ctgctgccac cgttgtttcc accgaagaaa 4380
ccaccgttgc cgtaaccacc acgacggttg ttgctaaaga agctgccacc gccacggcca 4440
ccgttgtagc cgccgttgtt gttattgtag ttgctcatgt tatttctggc acttcttggt 4500
tttcctctta agtgaggagg aacataacca ttctcgttgt tgtcgttgat gcttaaattt 4560
tgcacttgtt cgctcagttc agccataata tgaaatgctt ttcttgttgt tcttacggaa 4620
taccacttgc cacctatcac cacaactaac tttttcccgt tcctccatct cttttatatt 4680
ttttttctcg atcgagttca agagaaaaaa aaagaaaaag caaaaagaaa aaaggaaagc 4740
gcgcctcgtt cagaatgaca cgtatagaat gatgcattac cttgtcatct tcagtatcat 4800
actgttcgta tacatactta ctgacattca taggtataca tatatacaca tgtatatata 4860
tcgtatgctg cagctttaaa taatcggtgt cactacataa gaacaccttt ggtggaggga 4920

acatcgttgg taccattggg cgaggtggct tctcttatgg caaccgcaag agccttgaac 4980

gcactctcac tacggtgatg atcattcttg cctcgcagac aatcaacgtg gagggtaatt 5040

ctgctagcct ctgcaaagct ttcaagaaaa tgcgggatca tctcgcaaga gagatctcct 5100

actttctccc tttgcaaacc aagttcgaca actgcgtacg gcctgttcga aagatctacc 5160

accgctctgg aaagtgcctc atccaaaggc gcaaatcctg atccaaacct ttttactcca 5220

cgcgccagta gggcctcttt aaaagcttga ccgagagcaa tcccgcagtc ttcagtggtg 5280

tgatggtcgt ctatgtgtaa gtcaccaatg cactcaacga ttagcgacca gccggaatgc 5340

ttggccagag catgtatcat atggtccaga aaccctatac ctgtgtggac gttaatcact 5400

tgcgattgtg tggcctgttc tgctactgct tctgcctctt tttctgggaa gatcgagtgc 5460

tctatcgcta ggggaccacc ctttaaagag atcgcaatct gaatcttggt ttcatttgta 5520

atacgcttta ctagggcttt ctgctctgtc atctttgcct tcgtttatct tgcctgctca 5580

ttttttagta tattcttcga agaaatcaca ttactttata taatgtataa ttcattatgt 5640

gataatgcca atcgctaaga aaaaaaaaga gtcatccgct aggtggaaaa aaaaaaatga 5700

aaatcattac cgaggcataa aaaaatatag agtgtactag aggaggccaa gagtaataga 5760

aaaagaaaat tgcgggaaag gactgtgtta tgacttccct gactaatgcc gtgttcaaac 5820

gatacctggc agtgactcct agcgctcacc aagctcttaa aacgagaatt aagaaaaagt 5880

cgtcatcttt cgataagttt ttcccacagc aaagcaatag tagaaaaaaa caatgggaaa 5940

cgttgaatga agacaaagcg tcgtggttta aaaggaaata cgctcacgta catgctaggg 6000

aacaggaccg tgcagcggat cccgcgcatc aacaatattt tcacctgaat caggatattc 6060

ttctaatacc tgaatgctgt tttcccaccc caaccccaac cccaacccca accccaaccc 6120

caaccccaac cccaacccca accccaaccc caaccccaac cccaacccca accccaaccc 6180

caaccccaac cccaacccca accccaaccc caaccccaac cccaacccca accccaaccc 6240

caaccccaac cccaacccca accccaaccc caaccccaac cccaacccca accccaaccc 6300

caaccccaac cccaacccca accccaaccc caaccccaac cccaacccca accccaataa 6360

ttctttgaca ttgagtaaaa gttatttatt gatgtaatac tttgattttt agttatttat 6420

gaatgattaa gatgttaaaa tgtttaaatt ctataatatt ttgaatagtt tatatatgaa 6480

taaacataaa atattaatta aatctcaaaa atgactaagc tagctagtca aagattgaag 6540

ttcttatcaa aattatttta attaaaaatt ttatttttg tagttttcat aaatttttat 6600

aagtttttca ttcaaaaata aacttttaga agggtttttt gttataaaag atcaataaat 6660

aaaatgtttt tcaagaaatt aatgatgatt aataaatttt atctaaaaat gaagcttatc 6720

cctcgagggc tgcctcgcgc gtttcggtga tgacggtgaa aacctctgac acatgcagct 6780

cccggagacg gtcacagctt gtctgtaagc ggatgccggg agcagacaag cccgtcaggg 6840

cgcgtcagcg ggtgttggcg ggtgtcgggg cgcagccatg acccagtcac gtagcgatag 6900

cggagtgtat actggcttaa ctatgcggca tcagagcaga ttgtactgag agtgcaccat 6960

atgcggtgtg aaataccgca cagatgcgta aggagaaaat accgcatcag gcgctcttcc 7020

gcttcctcgc tcactgactc gctgcgctcg gtcgttcggc tgcggcgagc ggtatcagct 7080

cactcaaagg cggtaatacg gttatccaca gaatcagggg ataacgcagg aaagaacatg 7140

tgagcaaaag gccagcaaaa ggccaggaac cgtaaaaagg ccgcgttgct ggcgtttttc 7200

cataggctcc gccccctga cgagcatcac aaaaatcgac gctcaagtca gaggtggcga 7260

aacccgacag gactataaag ataccaggcg tttccccctg gaagctccct cgtgcgctct 7320

cctgttccga ccctgccgct taccggatac ctgtccgcct ttctcccttc gggaagcgtg 7380

gcgctttctc atagctcacg ctgtaggtat ctcagttcgg tgtaggtcgt tcgctccaag 7440

ctgggctgtg tgcacgaacc ccccgttcag cccgaccgct gcgccttatc cggtaactat 7500

cgtcttgagt ccaacccggt aagacacgac ttatcgccac tggcagcagc cactggtaac 7560

aggattagca gagcgaggta tgtaggcggt gctacagagt tcttgaagtg gtggcctaac 7620

tacggctaca ctagaaggac agtatttggt atctgcgctc tgctgaagcc agttaccttc 7680

ggaaaaagag ttggtagctc ttgatccggc aaacaaacca ccgctggtag cggtggtttt 7740

tttgtttgca agcagcagat tacgcgcaga aaaaaaggat ctcaagaaga tcctttgatc 7800

ttttctacgg ggtctgacgc tcagtggaac gaaaactcac gttaagggat tttggtcatg 7860

agattatcaa aaaggatctt cacctagatc cttttaaatt aaaaatgaag ttttaaatca 7920

atctaaagta tatatgagta aacttggtct gacagttacc aatgcttaat cagtgaggca 7980

cctatctcag cgatctgtct atttcgttca tccatagttg cctgactccc cgtcgtgtag 8040

ataactacga tacgggaggg cttaccatct ggccccagtg ctgcaatgat accgcgagac 8100

ccacgctcac cggctccaga tttatcagca ataaaccagc cagccggaag ggccgagcgc 8160

agaagtggtc ctgcaacttt atccgcctcc atccagtcta ttaattgttg ccgggaagct 8220

agagtaagta gttcgccagt taatagtttg cgcaacgttg ttgccattgc tgcaggcatc 8280

gtggtgtcac gctcgtcgtt tggtatggct tcattcagct ccggttccca acgatcaagg 8340

cgagttacat gatcccccat gttgtgcaaa aaagcggtta gctccttcgg tcctccgatc 8400

gttgtcagaa gtaagttggc cgcagtgtta tcactcatgg ttatggcagc actgcataat 8460

tctcttactg tcatgccatc cgtaagatgc ttttctgtga ctggtgagta ctcaaccaag 8520

tcattctgag aatagtgtat gcggcgaccg agttgctctt gcccggcgtc aacacgggat 8580

aataccgcgc cacatagcag aactttaaaa gtgctcatca ttggaaaacg ttcttcgggg 8640

cgaaaactct caaggatctt accgctgttg agatccagtt cgatgtaacc cactcgtgca 8700

cccaactgat cttcagcatc ttttactttc accagcgttt ctgggtgagc aaaaacagga 8760

aggcaaaatg ccgcaaaaaa gggaataagg gcgacacgga aatgttgaat actcatactc 8820

ttccttttc aatattattg aagcatttat cagggttatt gtctcatgag cggatacata 8880

tttgaatgta tttagaaaaa taaacaaata ggggttccgc gcacatttcc ccgaaaagtg 8940

ccacctgacg tctaagaaac cattattatc atgacattaa cctataaaaa taggcgtatc 9000

acgaggccct ttcgtcttca agaattaatt cggtcgaaaa aagaaaagga gagggccaag 9060

agggagggca ttggtgacta ttgagcacgt gagtatacgt gattaagcac acaaaggcag 9120

cttggagtat gtctgttatt aatttcacag gtagttctgg tccattggtg aaagtttgcg 9180

gcttgcagag cacagaggcc gcagaatgtg ctctagattc cgatgctgac ttgctgggta 9240

ttatatgtgt gcccaataga aagagaacaa ttgacccggt tattgcaagg aaaatttcaa 9300

gtcttgtaaa agcatataaa aatagttcag gcactccgaa atacttggtt ggcgtgtttc 9360

gtaatcaacc taaggaggat gttttggctc tggtcaatga ttacggcatt gatatcgtcc 9420

aactgcatgg agatgagtcg tggcaagaat accaagagtt cctcggtttg ccagttatta 9480

aaagactcgt atttccaaaa gactgcaaca tactactcag tgcagcttca cagaaacctc 9540

attcgtttat tcccttgttt gattcagaag caggtgggac aggtgaactt ttggattgga 9600

actcgatttc tgactgggtt ggaaggcaag agagccccga aagcttacat tttatgttag 9660

ctggtggact gacgccagaa aatgttggtg atgcgcttag attaaatggc gttattggtg 9720

ttgatgtaag cggaggtgtg gagacaaatg gtgtaaaaga ctctaacaaa atagcaaatt 9780

tcgtcaaaaa tgctaagaaa taggttatta ctgagtagta tttatttaag tattgtttgt 9840

gcacttgcct gcaggccttt tgaaaagcaa gcataaaaga tctaaacata aaatctgtaa 9900

aataacaaga tgtaaagata atgctaaatc atttggcttt ttgattgatt gtacaggaaa 9960

atatacatcg caggggggttg actttttacca tttcaccgca atggaatcaa acttgttgaa 10020

gagaatgttc acaggcgcat acgctacaat gacccgattc ttgctagcct tttctcggtc 10080

ttgcaaacaa ccgccggcag cttagtatat aaatacacat gtacatacct ctctccgtat 10140

cctcgtaatc attttcttgt atttatcgtc ttttcgctgt aaaaacttta tcacacttat 10200

ctcaaataca cttattaacc gcttttacta ttatcttcta cgctgacagt aatatcaaac 10260

agtgacacat attaaacaca gtggtttctt tgcataaaca ccatcagcct caagtcgtca 10320

agtaaagatt tcgtgttcat gcagatagat aacaatctat atgttgataa ttagcgttgc 10380

ctcatcaatg cgagatccgt ttaaccggac cctagtgcac ttaccccacg ttcggtccac 10440

tgtgtgccga acatgctcct tcactatttt aacatgtgga attaattcta aatcctcttt 10500

atatgatctg ccgatagata gttctaagtc attgaggttc atcaacaatt ggattttctg 10560

tttactcgac ttcaggtaaa tgaaatgaga tgatacttgc ttatctcata gttaactcta 10620

agaggtgata cttatttact gtaaaactgt gacgataaaa ccggaaggaa gaataagaaa 10680

actcgaactg atctataatg cctatttttct gtaaagagtt taagctatga aagcctcggc 10740

attttggccg ctcctaggta gtgctttttt tccaaggaca aaacagtttc tttttcttga 10800

gcaggtttta tgtttcggta atcataaaca ataaataaat tatttcattt atgtttaaaa 10860

ataaaaaata aaaaagtatt ttaaattttt aaaaaagttg attataagca tgtgaccttt 10920

tgcaagcaat taaattttgc aatttgtgat tttaggcaaa agttacaatt tctggctcgt 10980

gtaatatatg tatgctaaag tgaactttta caaagtcgat atggacttag tcaaaagaaa 11040

ttttcttaaa aatatatagc actagccaat ttagcacttc tttatgagat atattataga 11100

ctttattaag ccagatttgt gtattatatg tatttacccg gcgaatcatg gacatacatt 11160

ctgaaatagg taatattctc tatggtgaga cagcatagat aacctaggat acaagttaaa 11220

agctagtact gttttgcagt aatttttttc ttttttataa gaatgttacc acctaaataa 11280

gttataaagt caatagttaa gtttgatatt tgattgtaaa ataccgtaat atatttgcat 11340

gatcaaaagg ctcaatgttg actagccagc atgtcaacca ctatattgat caccgatata 11400

tggacttcca caccaactag taatatgaca ataaattcaa gatattcttc atgagaatgg 11460

cccaga 11466